# Prioritizing High-Risk Sub-Saharan African Adolescent Girls and Young Women for Prevention Interventions Using a Bayesian Spatial Model

Supplemental Content

Contact for technical questions: Steve Gutreuter
sgutreuter@cdc.gov

## Table of Contents



# 1 Supplemental Methods

## 1.1 Model definition

### 1.1.1 Accommodating the survey designs

The PHIA surveys follow stratified, multi-stage sampling designs. Strata are typically defined by first-level subnational areas ("regions"). The PHIA surveys from Cameroon and Kenya were additionally stratified by urban versus rural population density. The primary sampling units are census enumeration areas ("clusters") and were randomly selected, within strata, with probability proportional to the numbers of households or population.

Fixed numbers of households were selected within each cluster using systematic sampling with a random start. Additional details about all PHIA surveys can be found at `https://phia-data.icap.columbia.edu/`.

It is important to accommodate, to the extent possible, features of the survey design in Bayesian model-based estimation and prediction from survey data [1–4]. Inclusion of survey strata as a predictor strains parameter identification in the presence of the spatial error structures, and we therefore ignored the stratum identifiers in our models. Some of the PHIA surveys were also stratified by urban versus rural residency. We included the urbanicity indicator in our models.

Model-based inference from survey data predicts outcomes in sampled and non-sampled units by including the survey sampling probabilities or weights as predictors [1, 5, 6] and can outperform design-based estimation in terms of root mean-squared error [7]. However, our models are fitted to pooled data from 13 probability surveys. Rather than using the final blood sampling weights from each survey, we re-scaled those weights so that the country-specific sums of the re-scaled weights equaled the effective sample sizes for the surveys [8, 9]. The effective sample size for a survey was the actual sample size divided by the design effect for estimation of HIV prevalence. Therefore those re-scaled weights have effective sample size as their common basis across all surveys.

### 1.1.2 Likelihoods

Let $y_{mij} \in \{0, 1\}$, denote the absence/presence of HIV infection in young female $i = 1, \ldots n_y$ from area $j = 1, \ldots, n_a$ in country $m = 1, \ldots, 13$, and let $\boldsymbol{y} = (y_{mij})^{\mathsf{T}}$, where $\boldsymbol{a}^{\mathsf{T}}$ denotes vector/matrix transpose of $\boldsymbol{a}$. Let $\boldsymbol{p}$ denote $(p_{mij})^{\mathsf{T}}$, where the $p_{mij}$ are the probabilities of infection for young female $i$ in area $j$ of country $m$. Let $\boldsymbol{Z}$ denote a matrix of "fixed"-effect demographic and behavioral covariates having coefficient vector $\boldsymbol{\beta}_Z$. Let $\boldsymbol{x}$ represent the latent vector $(x_{mj})^{\mathsf{T}}$, the elements of which are the population viral loads in area $j$ of country $m$. $\boldsymbol{x}$ is observed indirectly through the proxy variable $\boldsymbol{w} = (w_{mjk})^{\mathsf{T}}$ where the $w_{mjk} = \log_{10}(\mathrm{VL}_{mjk} + 1), k = 1, \ldots, n_x$, and where $\mathrm{VL}_{mjk}$ is the viral load, measured in units of $\mathrm{copies} \cdot \mathrm{ml}^{-1}$ for individual $mjk$ among $n_x$ females and males of all ages in the corresponding areas. By definition, $\mathrm{VL}_{mjk} \equiv 0$ for HIV-negative individuals. The inclusion of $\boldsymbol{x}$ as a predictor of $\boldsymbol{y}$ requires a classical measurement error model [10] given by

$$\boldsymbol{y} \sim \mathrm{Bernoulli}(\boldsymbol{p}), \tag{1}$$

$$\mathrm{logit}(\boldsymbol{p}) = \beta_0 \mathbf{1}_y + \beta_x \boldsymbol{x} + \boldsymbol{Z}\boldsymbol{\beta}_Z + \boldsymbol{b}_Y + \boldsymbol{v}_c + \boldsymbol{v}_e + (\epsilon_{mjk})^{\mathsf{T}}, \tag{2}$$

$$\boldsymbol{w} = \boldsymbol{x} + (\epsilon_{Wmjk})^{\mathsf{T}}, \tag{3}$$

$$\boldsymbol{x} = \alpha_0 \mathbf{1}_x + \boldsymbol{b}_X + (\epsilon_{Xmj})^{\mathsf{T}} \tag{4}$$

where $\beta_0$ is an intercept in the linear predictor (eq. 2) and $\mathbf{1}_y$ denotes a vector of $n_y$ 1's. $\beta_x$ is a hyperparameter representing the logit-linear slope in $\boldsymbol{x}$. The *iid* random vectors $\boldsymbol{v}_c \sim \mathrm{N}(0, \tau_c)$ and $\boldsymbol{v}_e \sim \mathrm{N}(0, \tau_e)$ represent country- and enumeration-area-(cluster) level random effects having precisions $\tau_c$ and $\tau_e$, respectively, and the $\epsilon_{mjk}$ are individual-level $\mathrm{N}(0, \tau_y)$ random effects. The $\epsilon_{Wmjk}$ and $\epsilon_{Xmj}$ are independently and identically distributed (*iid*) Gaussian random errors having means 0 and precisions 10 and $\tau_X$, respectively. The rather large fixed precision for $\epsilon_{Wmjk}$ forces $\boldsymbol{x}$ to approximate $\boldsymbol{w}$. Equations 2 and 3 comprise the observation process and eq. 4 is a latent process.

This joint model contains a Bernoulli likelihood for $\boldsymbol{y}$ (eq. 1) and Gaussian likelihoods $\boldsymbol{w} \sim \mathrm{N}(\boldsymbol{x}, 10)$ (eq. 3) and $\boldsymbol{x} \sim \mathrm{N}(\alpha_0 \mathbf{1}_x + \boldsymbol{b}_X, \tau_x)$ (eq. 4). The coefficient $\alpha_0$ is an intercept in the model for $\boldsymbol{x}$ and $\mathbf{1}_x$ is vector of $m \times j$ 1's. The $\boldsymbol{b}_Y$ and $\boldsymbol{b}_X$ are spatially smoothed area-level random-effect vectors (see section 1.1.3, below).

The latent Gaussian random field is then given by $(\beta_0, \boldsymbol{x}^\intercal, \boldsymbol{\beta}_z^\intercal, \alpha_0)^\intercal$. Our primary interest is in estimates of $\boldsymbol{p}$, $\beta_x$, $\boldsymbol{x}$ and $\boldsymbol{\beta}_Z$.

### 1.1.3 Spatial smoothing

Honoring Tobler's first law of geography [11] that "everything is related to everything else, but near things are more related than distant things", we modeled spatial correlation in HIV status and PVL using the area-level BYM2 model [12,13]. The BYM2 model extends the more popular BYM model [14] by enabling scaling which facilitates hyperprior specification [15].

The BYM2 area-level random error vectors $\boldsymbol{b}$ have the form

$$\boldsymbol{b} = \frac{1}{\sqrt{\tau_b}} \left( \sqrt{1-\phi}\boldsymbol{v} + \sqrt{\phi}\boldsymbol{u}_* \right)$$

where $\boldsymbol{v} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$, $\boldsymbol{u}_* \sim \mathrm{N}(\mathbf{0}, \mathbf{Q}_*^-)$, $\mathbf{I}$ is the identity matrix, $\tau_b$ is a precision parameter. Herein all Gaussian distributions are parameterized using precision, which is the reciprocal of variance. The hyperparameter $\phi \in (0, 1)$ specifies the fraction of marginal standard error $1/\sqrt{\tau_b}$ explained by the scaled random effect $\boldsymbol{u}_*$, and $\boldsymbol{v}$ is an *iid* random effect, sometimes called the nugget. Note the the spatially independent nugget effect dominates as $\phi \to 1$ and the spatial component dominates as $\phi \to 0$. The matrix $\mathbf{Q}_*$ is a scaled version of the $mj \times mj$ spatial neighbor matrix $\mathbf{Q}$ having elements

$$Q_{gh} = \begin{cases} n_{\delta g} & \text{if } g = h, \\ -1 & \text{if } g \sim h, \\ 0 & \text{otherwise} \end{cases}$$

where $n_{\delta g}$ denotes the number of neighbors of area $g$, and $g \sim h$ denotes the condition that areas $g$ and $h$ are neighbors. The generalized variance of the random-effect vector $\boldsymbol{u}$ is given by

$$\sigma^2_{\mathrm{GV}}(\boldsymbol{u}) = \frac{1}{\tau} \exp\left( \frac{1}{n} \sum_{i=1}^{n} \log\left( \left[ \mathbf{Q}^- \right]_{ii} \right) \right).$$

Then, $\boldsymbol{u}_*$ is obtained by scaling $\boldsymbol{u}$ such that $\sigma^2_{\mathrm{GV}}(\boldsymbol{u}) = 1/\tau_b$, the marginal variance of $\boldsymbol{b}$ [13].

We imposed sum-to-zero constraints on all BYM2 structures. The graph for $\mathbf{Q}$ has disconnected components, including singletons (isolated unitary areas) because the data spans multiple, sometimes disconnected countries, and some countries include islands. If $\tau\mathbf{R}$ is the precision matrix of $\boldsymbol{v}$, then $\mathbf{R}$ is scaled so that the marginal variances of each connected component containing at least two areal units are 1, and singletons are given an $\mathrm{N}(0, 1)$ distribution.

### 1.1.4 Priors and hyperpriors

HIV infection is rare. We assigned vague independent $\mathrm{N}(0, 1/9)$ priors to $\beta_0$, the components of $\boldsymbol{\beta}_Z$ and $\alpha_0$, which convey almost no information on the logit scale. The likelihood for $\boldsymbol{x}$ contains two *iid* Gaussian components, $\epsilon_X$ and $\boldsymbol{v}$. Any practical effect of $\epsilon_{Xj}$ is minimized by assigning fixed precision $\tau_X = 10$. A moderately informative hyperprior is required for $\beta_x$. Based on preliminary exploratory plots of survey domain estimates, we anticipated that the 20th and 80th percentiles of $\beta_x$ might be approximately 4 and 7, respectively, so we chose for it a $\mathrm{N}(5, 1/4)$ hyperprior. We chose a vague Gamma(1, 0.01) hyperprior for $\tau_W$, which gives 0.025 and 0.95 percentiles of 0.056 and 4.026 for the standard deviation [6]. Finally, we chose penalized complexity (PC) hyperpriors [12] for the BYM2 structures $\boldsymbol{b}_Y$ and $\boldsymbol{b}_X$. The PC prior for the mixing parameter $\phi$ will automatically shrink towards 0 (no spatial smoothing) in the absence of evidence in the data. For both, we assigned PC priors for the mixing parameters $\phi$ such that $\Pr(\phi < 0.5) = 2/3$, which slightly favors simpler *iid* area-level random effects. The PC priors for the area-level precisions $\tau_b$ were chosen such that $\Pr(\mathrm{SD} > 0.2) = 0.1$.

## 1.2 Computation

R version 4.3.1 [16] was used for all computations. We approximated the joint posterior distributions of the latent random field using the INLA [17] package version INLA_2023-09-09. INLA uses computationally fast nested Laplace approximations and numerical integration. Computational speed is critical to our application because of the large number of survey observations and the high dimension of the latent field, for which Markov Chain or Hamiltonian Monte Carlo sampling would have been impractical.

### 1.2.1 Model variations

**Supplemental Table 1.** Joint model variations.

| | iid random effects | | |
|---|---|---|---|
| Model name | Country | Cluster | Weight covariate |
| M000 | No | No | No |
| M001 | No | No | Yes |
| M010 | No | Yes | No |
| M011 | No | Yes | Yes |
| M100 | Yes | No | No |
| M101 | Yes | No | Yes |
| M110 | Yes | Yes | No |
| M111 | Yes | Yes | Yes |

We fitted model variations ignoring the re-scaled weights, and also followed [6] by including the re-scaled weights in *B*-spline basis functions [18] in the linear predictor (eq. 2). We chose *B*-splines having 3 df. We fitted eight model variations excluding and including country-level iid random effects $\boldsymbol{v}_c$, cluster-level iid random effects $\boldsymbol{v}_e$ and *B*-spline basis functions of the re-scaled weights (Supplemental Table 1).

**Supplemental Table 2.** Descriptions of variables used in R code.

| Variable | Type | Likelihood | Description |
|---|---|---|---|
| Y1 | binary | 1 | Response matrix column 1 (Bernoulli likelihood; HIV status) |
| Y2 | integer | 2 | Response matrix column 2 ('zeros' data stack; See https://doi.org/10.1111/rssc.12069) |
| Y3 | numeric | 3 | Response matrix column 3 (2nd Gaussian likelihood; log10VL) |
| beta.0 | integer | 1 | Column of 1's for intercept |
| alpha.0 | integer | 1 | Column of 1's for intercept |
| beta.x | integer | 1 | Column containing target snu numbers |
| idx.x | integer | 3 | Index for target subnational estimation units in spatial neighbors matrix (1-1 |
| idx.y | integer | 1 | Index for target subnational estimation units in spatial neighbors matrix |
| idx.ea.x | integer | 3 | Index for clusters (aka centroids or enumeration areas (1-6 |
| idx.ea.y | integer | 1 | Index for clusters (aka centroids or enumeration areas (1-6 |
| wgt.x | integer | 2 | Linear predictor formulation (see https://doi.org/10.1111/rssc.12069) |
| idx.iso3.x | integer | 3 | Index for iso3 country codes (1-13) |
| idx.iso3.y | integer | 1 | Index for iso3 country codes (1-13) |
| Ntrials | integer | 1 | Column of 1's for number of trials to obtain a Bernoulli likelihood |
| schPri | binary | 1 | Indicator for completion of primary school (baseline no/unknown school) |
| schSec | binary | 1 | Indicator for completion of secondary school (baseline no/unknown school) |
| schPSec | binary | 1 | Indicator for completion of more than secondary school (baseline no/unknown school) |
| txsex | binary | 1 | Indicator for transactional/commercial sex in past yr (baseline no) |
| debut1_age0_workyr0 | binary | 1 | Indicator for sexual debut > 16 and age 15-19/no work in past year (baseline never had sex/age 15-19/no work in past yr) |
| debut2_age0_workyr0 | binary | 1 | Indicator for sexual debut $\leq$ 16 and age 15-19/no work in past year (baseline never had sex/age15-19/no work) |
| debut0_age1 | binary | 1 | Indicator for never had sex and age 20-24 (baseline never had sex and age 15-19/no work) |
| debut1_age1 | binary | 1 | Indicator for debut > 16 and age 20-24 (baseline never had sex and age 15-19/no work) |
| debut2_age1 | binary | 1 | Indicator for debut $\leq$ 16 and age 20-24 (baseline never had sex and age 15-19/no work) |
| debut1_workyr1 | binary | 1 | Indicator debut > 16 and worked during past year (baseline never had sex and age 15-19/no work) |
| debut2_workyr1 | binary | 1 | Indicator debut $\leq$ 16 and worked during past year (baseline never had sex and age 15-19/no work) |
| preg0_partN | binary | 1 | Indicator for partner HIV-negative and no history of pregnancy (baseline no partner and no history of pregnancy) |
| preg0_partU | binary | 1 | Indicator for partner HIV-unknown and no pregnancy (baseline no partner and no history of pregnancy) |
| preg0_partP | binary | 1 | Indicator for partner HIV-positive and no pregnancy (baseline no partner and no history of pregnancy) |
| preg1_part0 | binary | 1 | Indicator for no partner and history of pregnancy (baseline no partner and no history of pregnancy) |
| preg1_partN | binary | 1 | Indicator for partner HIV-negative and history of pregnancy (baseline no partner and no history of pregnancy) |
| preg1_partU | binary | 1 | Indicator for partner HIV-unknown and history of pregnancy (baseline no partner and no history of pregnancy) |
| preg1_partP | binary | 1 | Indicator for partner HIV-positive and history of pregnancy (baseline no partner and no history of pregnancy) |
| wt_scaled.y | numeric | 1 | Rescaled survey weights for HIV status |

### 1.2.2 Description of variables and data preparation for INLA

The descriptions of the variables used in the R code, below, are given in Supplemental Table 2. Recall that the joint spatial model includes three likelihoods. The column "Likelihood" refers to the likelihood for which each variable informs.

As of this writing, the handling of factor covariates by INLA was unreliable, sometimes leading to fully saturated parameterizations. Therefore we created binary indicators for the levels of all factor variables. INLA also has specific requirements for data formats for joint measurement error models [10]. The model contains three likelihoods, and there for the response matrix $\mathbf{Y}$ must contain as many columns. The R code for both tasks follows.

```
library(tidyverse, quietly = TRUE)
################################################################################
## Define file paths
################################################################################
here::i_am("DREAMS_models/code/BYM2_AFR.R")
workpath <- here::here("DREAMS_models/code")
datapath <- here::here("data")
outpath <- here::here("output")
setwd(workpath)
############################3#####################################################
## Get the viral load data from all respondents and create tsnu_num, which is
## the unique tsnu (target subnational unit) number for each observation
################################################################################
vldat <- readRDS(file.path(datapath, "PHIA_VL+geo_data.rds"))
vldat <- vldat[complete.cases(vldat[, c("tsnu", "log10VL")]), ]
vldat <- vldat %>%
    select(-tsnu_id)
snun <- unique(vldat$tsnu)
snu <- data.frame(tsnu = snun,
                  tsnu_num = seq_along(along.with = snun))
vldat <- left_join(vldat, snu, by = "tsnu")
vldat <- vldat %>%
    arrange(tsnu, centroidid, personid)
################################################################################
## Create indexes for the country iso3 identifiers
################################################################################
iso3 <- unique(vldat$iso3)
isoidx <- data.frame(iso3 = iso3, idx.iso3 = 1:length(iso3))
vldat <- vldat %>%
    left_join(isoidx, by = "iso3")
################################################################################
## Create indexes for the centroids (survey enumeration areas)
################################################################################
centroidid <- unique(vldat$centroidid)
clstridx <- data.frame(centroidid = centroidid, idx.ea = 1:length(centroidid))
vldat <- vldat %>%
    left_join(clstridx, by = "centroidid") %>%
    arrange(tsnu, centroidid, personid)
################################################################################
## Get the screening data from the AGYW, retaining complete cases on the key
## variables, and code the interpretable forms of terms having interactions
################################################################################
qdat <- readRDS(file.path(datapath, "PHIA_DREAMS_Data_recoded.rds"))
qdat <- qdat %>%
    dplyr::select(iso3, centroidid, personid, area_order, PHIA_strat_nm,
                  urban, age_cat, debut, workyr, pregever, txsex,
                  school, partHIV, HIVstat,
                  tsnu, wt_phat_scaled) %>%
```

```
mutate(age_cat = if_else(age_cat == "15-19", 0L, 1L),
       pregever = if_else(pregever == "No", 0L, 1L),
       txsex = if_else(txsex == "No transactional sex in past yr", 0L, 1L),
       workyr = if_else(workyr == "No", 0L, 1L),
       age0_partN = if_else(age_cat == 0L & partHIV == "Partner negative",
                            1L, 0L),
       age0_partU  = if_else(age_cat == 0L & partHIV == "Partner status unknown",
                             1L, 0L),
       age0_partP = if_else(age_cat == 0L & partHIV == "Partner positive",
                            1L, 0L),
       age1_part0 = if_else(age_cat == 1L & partHIV == "No sex partner",
                            1L, 0L),
       age1_partN = if_else(age_cat == 1L & partHIV == "Partner negative",
                            1L, 0L),
       age1_partU = if_else(age_cat == 1L & partHIV == "Partner status unknown",
                            1L, 0L),
       age1_partP = if_else(age_cat == 1L & partHIV == "Partner positive",
                            1L, 0L),
       debut1_age0 = if_else(age_cat == 0L &
                             debut == ">16", 1L, 0L),
       debut2_age0 = if_else(age_cat ==0L &
                             debut == "<=16", 1L, 0L),
       debut0_age1 = if_else(age_cat == 1L &
                             debut == "Never had sex", 1L, 0L),
       debut1_age1 = if_else(age_cat == 1L &
                             debut == ">16", 1L, 0L),
       debut2_age1= if_else(age_cat == 1L &
                            debut == "<=16", 1L, 0L),
       debut0_workyr1 = if_else(debut == "Never had sex" & workyr == 1L,
                                1L, 0L),
       debut1_workyr0 = if_else(debut == ">16" & workyr == 0L, 1L, 0L),
       debut1_workyr1 = if_else(debut == ">16" & workyr == 1L, 1L, 0L),
       debut2_workyr0 = if_else(debut == "<=16" & workyr == 0L, 1L, 0L),
       debut2_workyr1 = if_else(debut == "<=16" & workyr == 1L, 1L, 0L),
       schP_part0 = if_else(school == "Primary" &
                            partHIV == "No sex partner", 1L, 0L),
       schS_part0 = if_else(school == "Secondary" &
                            partHIV == "No sex partner", 1L, 0L),
       schPS_part0 = if_else(school == "Post-secondary" &
                             partHIV == "No sex partner", 1L, 0L),
       sch0_partN = if_else(school == "Never attended/unknown" &
                            partHIV == "Partner negative", 1L, 0L),
       schP_partN = if_else(school == "Primary" &
                            partHIV == "Partner negative", 1L, 0L),
       schS_partN = if_else(school == "Secondary" &
                            partHIV == "Partner negative", 1L, 0L),
       schPS_partN = if_else(school == "Post-secondary" &
                             partHIV == "Partner negative", 1L, 0L),
```

```
            sch0_partU = if_else(school == "Never attended/unknown" &
                                 partHIV == "Partner status unknown", 1L, 0L),
            schP_partU = if_else(school == "Primary" &
                                 partHIV == "Partner status unknown", 1L, 0L),
            schS_partU = if_else(school == "Secondary" &
                                 partHIV == "Partner status unknown", 1L, 0L),
            schPS_partU = if_else(school == "Post-secondary" &
                                  partHIV == "Partner status unknown", 1L, 0L),
            sch0_partP = if_else(school == "Never attended/unknown" &
                                 partHIV == "Partner positive", 1L, 0L),
            schP_partP = if_else(school == "Primary" &
                                 partHIV == "Partner positive", 1L, 0L),
            schS_partP = if_else(school == "Secondary" &
                                 partHIV == "Partner positive", 1L, 0L),
            schPS_partP = if_else(school == "Post-secondary" &
                                  partHIV == "Partner positive", 1L, 0L),
            preg0_partN = if_else(pregever == 0L &
                                  partHIV == "Partner negative", 1L, 0L),
            preg0_partU = if_else(pregever == 0L &
                                  partHIV == "Partner status unknown", 1L, 0L),
            preg0_partP = if_else(pregever == 0L &
                                  partHIV == "Partner positive", 1L, 0L),
            preg1_part0 = if_else(pregever == 1L & partHIV == "No sex partner",
                                  1L, 0L),
            preg1_partN = if_else(pregever == 1L & partHIV == "Partner negative",
                                  1L, 0L),
            preg1_partU = if_else(pregever == 1L & partHIV == "Partner status unknown",
                                  1L, 0L),
            preg1_partP = if_else(pregever == 1L & partHIV == "Partner positive",
                                  1L, 0L)
            )
qdat <- qdat[complete.cases(qdat), ]
qdat <- left_join(qdat, snu, by = "tsnu")
qdat <- qdat %>%
    arrange(tsnu, centroidid, personid)
set.seed(98475)
Nfolds <- 10
qdat$fold <- sample(1:Nfolds, size = nrow(qdat), replace = TRUE)
rm(snu, snun)
################################################################################
## Get country and cluster identifiers from vldat
################################################################################
indxs <- vldat %>%
    select(tsnu, centroidid, personid, idx.iso3, idx.ea)
qdat <- qdat %>%
    left_join(indxs, by = c("tsnu", "centroidid", "personid"))
################################################################################
## Expand factors to indicator vectors to overcome INLA's inconsistent handling
```

```
## of factors, and code interactions.  As of the time of this work, it was not
## safe to use factors as covariates with INLA because they would sometimes
## yield over-parameterized models.
################################################################################
(n.Y <- nrow(qdat))
(n.W <- nrow(vldat))
age_cat <- c(qdat$age_cat, rep(NA, n.Y), rep(NA, n.W))
## partHIV; baseline is no partner
fmm1 <- model.matrix(~ partHIV, data = qdat)[, -1]
partNeg <- c(fmm1[, 1], rep(NA, n.Y), rep(NA, n.W))     ## Partner negative
partUnk <- c(fmm1[, 2], rep(NA, n.Y), rep(NA, n.W))     ## Unknown parter status
partPos <- c(fmm1[, 3], rep(NA, n.Y), rep(NA, n.W))     ## Partner positive
age1xpartN <- age_cat * partNeg
age1xpartU <- age_cat * partUnk
age1xpartP <- age_cat * partPos
## school; baseline is no school
fmm2 <- model.matrix(~ school - 1, data = qdat)[, -1]
schPri <- c(fmm2[, 1], rep(NA, n.Y), rep(NA, n.W))      ## Primary only
schSec <- c(fmm2[, 2], rep(NA, n.Y), rep(NA, n.W))      ## Secondary
schPSec <- c(fmm2[, 3], rep(NA, n.Y), rep(NA, n.W))     ## Post-secondary
schPxpN <- schPri * partNeg
schPxpU <- schPri * partUnk
schPxpP <- schPri * partPos
schSxpN <- schSec * partNeg
schSxpU <- schSec * partUnk
schSxpP <- schSec * partPos
schPSxpN <- schPSec * partNeg
schPSxpU <- schPSec * partUnk
schPSxpP <- schPSec * partPos
## debut; baseline is never had sex
fmm3 <- model.matrix(~ as.factor(debut), data = qdat)[, -1]
debutgt16 <- c(fmm3[, 1], rep(NA, n.Y), rep(NA, n.W)) ## Debut > 1
debutle16 <- c(fmm3[, 2], rep(NA, n.Y), rep(NA, n.W)) ## Debut <= 16
age1xdebut1 <- age_cat * debutgt16
age1xdebut2 <- age_cat * debutle16
rm(fmm1, fmm2, fmm3)
################################################################################
## Assemble the INLA-structured data for join measurement error modeling.  Note
## that the matrix of outcomes Y has three columns (one for each likelihood).
## See the Supplement to Muff et al. 2015 (https://doi.org/10.1111/rssc.12069)
## and/or the Supplement to section 3.2 in Krainski et al. 2021
## (https://becarioprecario.bitbucket.io/spde-gitbook/index.html) for an
## explanation of the rationale behind the INLA data stack Y for joint
## measurement-error models.
################################################################################
Y <- matrix(rep(NA, (3 * (2 * n.Y + n.W))), ncol = 3)
Y[1:n.Y, 1] <- qdat$HIVstat                   ## col 1: response variable
Y[(n.Y + (1:n.Y)), 2] <- rep(0, n.Y)          ## col 2: "Zeros" stack
```

```
Y[(2 * n.Y + (1:n.W)), 3] <- vldat$log10VL    ## col 3: observed surrogate for x
beta.0 <- c(rep(1, n.Y), rep(NA, n.Y), rep(NA, n.W))
alpha.0 <- c(rep(NA, n.Y), rep(1, n.Y), rep(NA, n.W))
beta.x <- c(qdat$tsnu_num, rep(NA, n.Y), rep(NA, n.W))
idx.x <- c(rep(NA, n.Y), qdat$tsnu_num, vldat$tsnu_num)
idx.y <- c(qdat$tsnu_num, rep(NA, n.Y), rep(NA, n.W))
idx.ea.x <- c(rep(NA, n.Y), rep(NA, n.Y), vldat$idx.ea)
idx.ea.y <- c(qdat$idx.ea, rep(NA, n.Y), rep(NA, n.W))
idx.iso3.x <- c(rep(NA, n.Y), rep(NA, n.Y), vldat$idx.iso3)
idx.iso3.y <- c(qdat$idx.iso3, rep(NA, n.Y), rep(NA, n.W))
wgt.x <- c(rep(1, n.Y), rep(-1, n.Y), rep(1, n.W))
iso3 <- c(qdat$iso3, rep(NA, n.Y), vldat$iso3)
partHIV <- c(qdat$partHIV, rep(NA, n.Y), rep(NA, n.W))
workyr <- c(qdat$workyr, rep(NA, n.Y), rep(NA, n.W))
debutgt16 <- debutgt16
debutle16 <- debutle16
debut1xworkyr <- debutgt16 * workyr
debut2xworkyr <- debutle16 * workyr
pregever <- c(qdat$pregever, rep(NA, n.Y), rep(NA, n.W))
partNxpreg <- partNeg * pregever
partUxpreg <- partUnk * pregever
partPxpreg <- partPos * pregever
txsex <- c(qdat$txsex, rep(NA, n.Y), rep(NA, n.W))
Ntrials <- c(rep(1, n.Y), rep(NA, n.Y), rep(NA, n.W))
urban.x <- c(rep(NA, n.Y), rep(NA, n.Y), vldat$urban)
urban.y <- c(qdat$urban, rep(NA, n.Y), rep(NA, n.W))
wt_scaled.x <- c(rep(NA, n.Y), rep(NA, n.Y), vldat$wt_PVL_scaled)
wt_scaled.y <- c(qdat$wt_phat_scaled, rep(NA, n.Y), rep(NA, n.W))
fold <- c(qdat$fold, rep(NA, n.Y), rep(NA, n.W))
tsnu <- c(qdat$tsnu, rep(NA, n.Y), vldat$tsnu)
area_order <- c(qdat$area_order, rep(NA, n.Y), vldat$area_order)
PHIA_strat_nm <- c(qdat$PHIA_strat_nm, rep(NA, n.Y), vldat$PHIA_strat_nm)
centroidid <- c(qdat$centroidid, rep(NA, n.Y), vldat$centroidid)
personid <- c(qdat$personid, rep(NA, n.Y), vldat$personid)
data_AFR <- data.frame(Y1 = Y[, 1], Y2 = Y[, 2], Y3 = Y[, 3],
                       beta.0 = beta.0,
                       beta.x = beta.x,
                       idx.x = idx.x,
                       idx.y = idx.y,
                       idx.iso3.x = idx.iso3.x,
                       idx.iso3.y = idx.iso3.y,
                       idx.ea.x = idx.ea.x,
                       idx.ea.y = idx.ea.y,
                       alpha.0 = alpha.0,
                       wgt.x = wgt.x,
                       Ntrials = Ntrials,
                       wt_scaled.x = wt_scaled.x,
                       wt_scaled.y = wt_scaled.y,
```

```
fold = fold,
iso3 = iso3,
centroidid = centroidid,
personid = personid,
PHIA_strat_nm = PHIA_strat_nm,
tsnu = tsnu,
area_order = area_order,
partHIV = partHIV,
debutgt16 = debutgt16,
debutle16 = debutle16,
schPri = schPri,
schSec = schSec,
schPSec = schPSec,
partNeg = partNeg,
partUnk = partUnk,
partPos = partPos,
age_cat = age_cat,
workyr = workyr,
pregever = pregever,
txsex = txsex,
urban.x = urban.x,
urban.y = urban.y,
partNxpreg = partNxpreg,
partUxpreg = partUxpreg,
partPxpreg = partPxpreg,
age1xpartN = age1xpartN,
age1xpartU = age1xpartU,
age1xpartP = age1xpartP,
age1xdebut1 = age1xdebut1,
age1xdebut2 = age1xdebut2,
debut1xworkyr = debut1xworkyr,
debut2xworkyr = debut2xworkyr,
schPxpN = schPxpN,
schPxpU = schPxpU,
schPxpP = schPxpP,
schSxpN = schSxpN,
schSxpU = schSxpU,
schSxpP = schSxpP,
schPSxpN = schPSxpN,
schPSxpU = schPSxpU,
schPSxpP = schPSxpP,
age0_partN = c(qdat$age0_partN, rep(NA, n.Y),
               rep(NA, n.W)),
age0_partU = c(qdat$age0_partU, rep(NA, n.Y),
               rep(NA, n.W)),
age0_partP = c(qdat$age0_partP, rep(NA, n.Y),
               rep(NA, n.W)),
age1_part0 = c(qdat$age1_part0, rep(NA, n.Y),
```

```
                rep(NA, n.W)),
age1_partN = c(qdat$age1_partN, rep(NA, n.Y),
                rep(NA, n.W)),
age1_partU = c(qdat$age1_partU, rep(NA, n.Y),
                rep(NA, n.W)),
age1_partP = c(qdat$age1_partP, rep(NA, n.Y),
                rep(NA, n.W)),
debut1_age0 = c(qdat$debut1_age0,  rep(NA, n.Y),
                rep(NA, n.W)),
debut2_age0 = c(qdat$debut2_age0, rep(NA, n.Y),
                 rep(NA, n.W)),
debut0_age1 = c(qdat$debut0_age1, rep(NA, n.Y),
                 rep(NA, n.W)),
debut1_age1 = c(qdat$debut1_age1, rep(NA, n.Y),
                 rep(NA, n.W)),
debut2_age1 = c(qdat$debut2_age1, rep(NA, n.Y),
                 rep(NA, n.W)),
debut0_workyr1 = c(qdat$debut0_workyr1, rep(NA, n.Y),
                    rep(NA, n.W)),
debut1_workyr0 = c(qdat$debut1_workyr0, rep(NA, n.Y),
                    rep(NA, n.W)),
debut1_workyr1 = c(qdat$debut1_workyr1, rep(NA, n.Y),
                    rep(NA, n.W)),
debut2_workyr0 = c(qdat$debut2_workyr0, rep(NA, n.Y),
                    rep(NA, n.W)),
debut2_workyr1 = c(qdat$debut2_workyr1, rep(NA, n.Y),
                    rep(NA, n.W)),
schP_part0 = c(qdat$schP_part0, rep(NA, n.Y),
                rep(NA, n.W)),
schS_part0 = c(qdat$schS_part0, rep(NA, n.Y),
                rep(NA, n.W)),
schPS_part0 = c(qdat$schPS_part0, rep(NA, n.Y),
                 rep(NA, n.W)),
sch0_partN = c(qdat$sch0_partN, rep(NA, n.Y),
                rep(NA, n.W)),
schP_partN = c(qdat$schP_partN, rep(NA, n.Y),
                rep(NA, n.W)),
schS_partN = c(qdat$schS_partN, rep(NA, n.Y),
                rep(NA, n.W)),
schPS_partN = c(qdat$schPS_partN, rep(NA, n.Y),
                 rep(NA, n.W)),
sch0_partU = c(qdat$sch0_partU, rep(NA, n.Y),
                rep(NA, n.W)),
schP_partU = c(qdat$schP_partU, rep(NA, n.Y),
                rep(NA, n.W)),
schS_partU = c(qdat$schS_partU, rep(NA, n.Y),
                rep(NA, n.W)),
schPS_partU = c(qdat$schPS_partU, rep(NA, n.Y),
```

```
                  rep(NA, n.W)),
sch0_partP = c(qdat$sch0_partP, rep(NA, n.Y),
                  rep(NA, n.W)),
schP_partP = c(qdat$schP_partP, rep(NA, n.Y),
                  rep(NA, n.W)),
schS_partP = c(qdat$schS_partP, rep(NA, n.Y),
                  rep(NA, n.W)),
schPS_partP = c(qdat$schPS_partP, rep(NA, n.Y),
                  rep(NA, n.W)),
preg0_partN = c(qdat$preg0_partN, rep(NA, n.Y),
                  rep(NA, n.W)),
preg0_partU = c(qdat$preg0_partU, rep(NA, n.Y),
                  rep(NA, n.W)),
preg0_partP = c(qdat$preg0_partP, rep(NA, n.Y),
                  rep(NA, n.W)),
preg1_part0 = c(qdat$preg1_part0, rep(NA, n.Y),
                  rep(NA, n.W)),
preg1_partN = c(qdat$preg1_partN, rep(NA, n.Y),
                  rep(NA, n.W)),
preg1_partU = c(qdat$preg1_partU, rep(NA, n.Y),
                  rep(NA, n.W)),
preg1_partP = c(qdat$preg1_partP, rep(NA, n.Y),
                  rep(NA, n.W)))
```

### 1.2.3 Example model code

The model variation (M111) which included country- and area-level random effects, and also a *B*-spline basis function of the rescaled survey weights having 3 df in the Bernoulli likelihood was fitted using the following code, below. Other model variations include or excluded the other combinations of those effects. The country-level random effects are included using `f(idx.iso3.y,...)`, the area-level random effects are included using `f(idx.ea.y, ...)` and the smoothed weight covariate is included as `bs(wt_scaled.y, df =3)` in the following R code example:

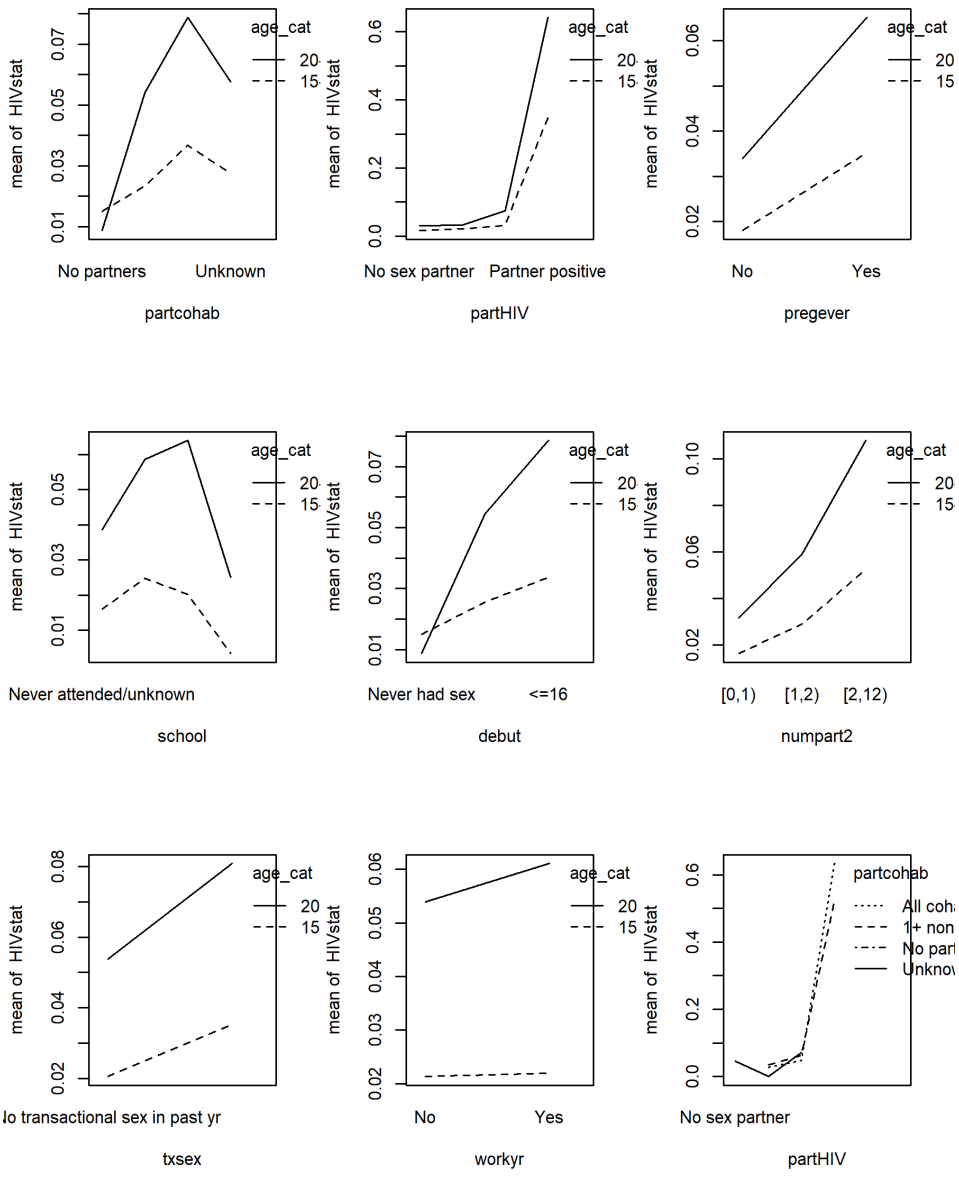
```
library(tidyverse, quietly = TRUE)
library(INLA)
library(brinla)
library(splines)
here::i_am("DREAMS_models/code/BYM2_AFR.R")
workpath <- here::here("DREAMS_models/code")
datapath <- here::here("data")
outpath <- here::here("output")
setwd(workpath)
################################################################################
## Get the spatial graph
################################################################################
g <- file.path(datapath, "NB_Graph_AF.dat")
```

```
################################################################################
## Get the INLA-structured joint measurement error model data for all PHIA
## countries
################################################################################
AFRdata <- readRDS(file = file.path(datapath, "BYM2_mem_AFR_data.rds"))
Y <- as.matrix(AFRdata[c("Y1", "Y2", "Y3")])
n.Y <- attr(AFRdata, "nAGYW")
waic <- data.frame(NULL)
################################################################################
## Prior and hyperprior parameters
################################################################################
prior.beta.x = c(5, 1/4)   ## Mildly informative N(5, tau = 1/4) prior on beta.x
## Prior precision for 1st Gaussian likelihood (\epsilon_{Wkj})
prec.w <- 10                           ##  High precision fixed in 2nd likelihood
## Prior precision for 2nd Gaussian likelihood
prior.prec.x <- c(1, 0.01)    ## See Vandendijck et al 2016, Sec 3.1. Gives 0.95
prec.x <- 4                   ## range on sigma of (0.056, 4.036).
form111 <- Y ~ beta.0 - 1 + alpha.0 + urban.y +
    debut1_age0_workyr0 + debut2_age0_workyr0 + debut0_age1 + debut1_age1 +
    debut2_age1 +
    debut0_workyr1 + debut1_workyr1 + debut2_workyr1 +
    preg0_partN + preg0_partU + preg0_partP + preg1_part0 + preg1_partN +
    preg1_partU + preg1_partP +
    schPri + schSec + schPSec + txsex +
    bs(wt_scaled.y, df = 3) +
    f(idx.iso3.y, model = "iid",
      hyper = list(prec = list(prior = "pc.prec", param = c(1, 0.01)))) +
    f(idx.ea.y, model = "iid",
      hyper = list(prec = list(prior = "pc.prec", param = c(1, 0.01)))) +
    f(idx.x, wgt.x, model = "bym2", graph = g, scale.model = TRUE,
      adjust.for.con.comp = TRUE,
      hyper = list(phi = list(prior = "pc",
                              param = c(0.5, 2/3)),       ## Pr(phi < 0.5) = 2/3
                   prec = list(prior = "pc.prec",
                               param = c(0.2, 0.1)))) +    ## Pr(sd > 0.2) = 0.1
    f(beta.x, copy = "idx.x", hyper = list(beta = list(param = prior.beta.x,
                                                       fixed = FALSE))) +
    f(idx.y, model = "bym2", graph = g, scale.model = TRUE,
      adjust.for.con.comp = TRUE,
      hyper = list(phi = list(prior = "pc",
                              param = c(0.5, 2/3)),       ## Pr(phi < 0.5) = 2/3
                   prec = list(prior = "pc.prec",
                               param = c(0.2, 0.1))))      ## Pr(sd > 0.2) = 0.1
M111 <- inla(form111, Ntrials = Ntrials, data = AFRdata,
          family = c("binomial", "gaussian", "gaussian"),
          control.family = list(
            list(hyper = list()),                                ## Binomial
            list(hyper = list(                                   ## 1st Gaussian
```

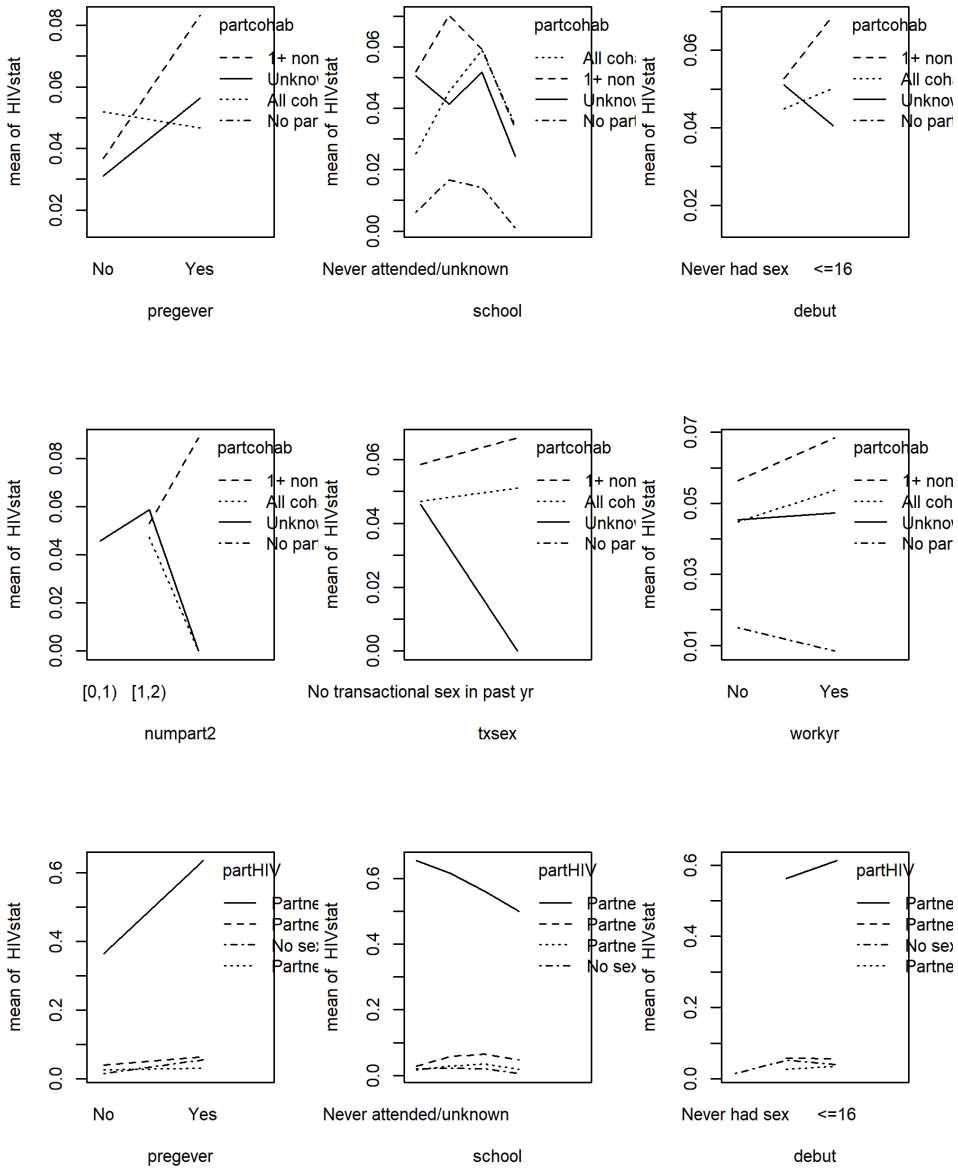
```
    prec = list(initial = log(prec.w),
                fixed = TRUE))),
  list(hyper = list(                                    ## 2nd Gaussian
    prec = list(initial=log(prec.x),
                param = prior.prec.x,
                fixed = FALSE)))),
control.fixed = list(
    mean = list(beta.0 = 0, beta = 0, alpha.0 = 0), ## Vague N(0, 1/9)
    prec = list(beta.0 = 1/9, beta = 1/9, alpha.0 = 1/9)),   ## priors
control.compute = list(config = TRUE, cpo = TRUE, waic = TRUE))
```

# 2 Supplemental Results

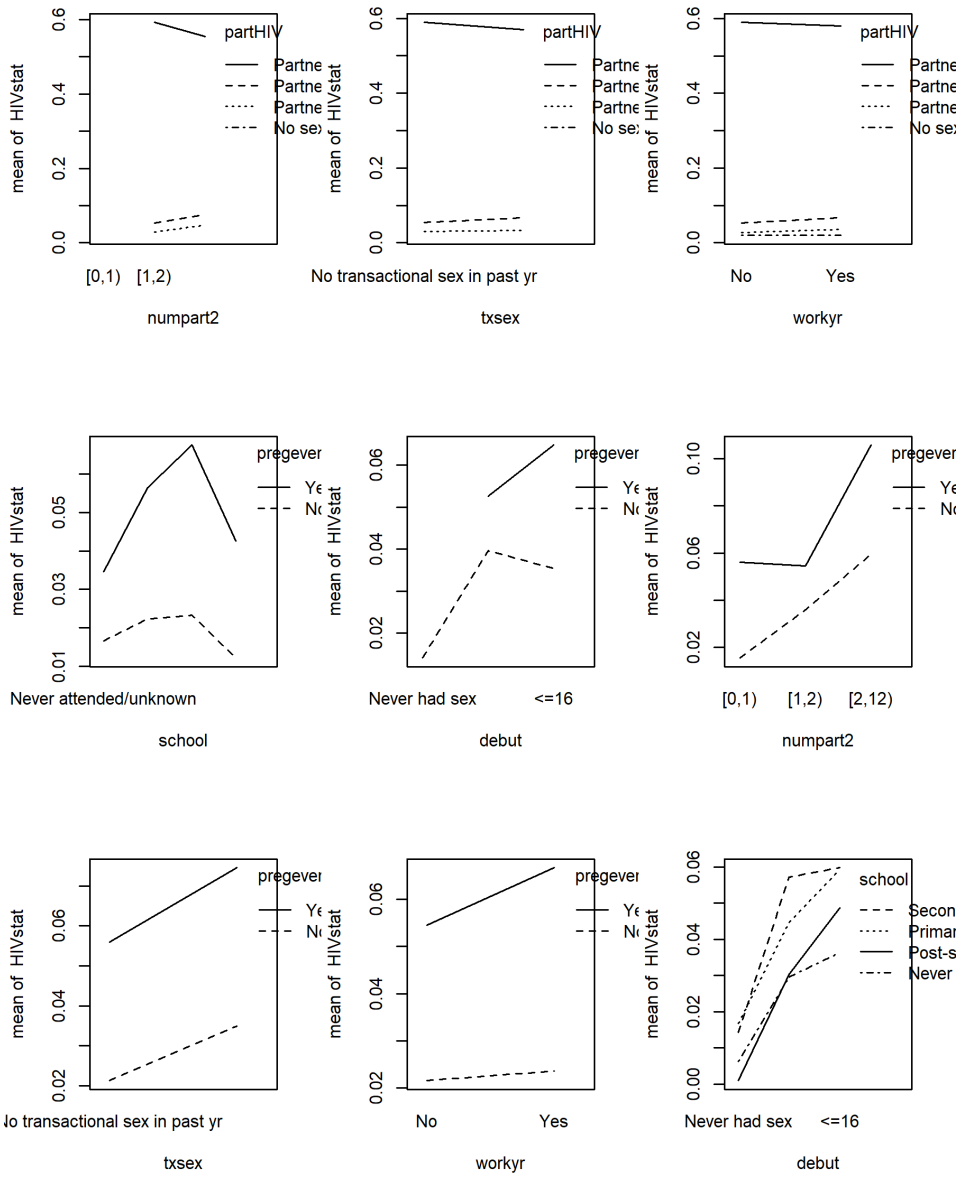
## 2.1 Interaction plots



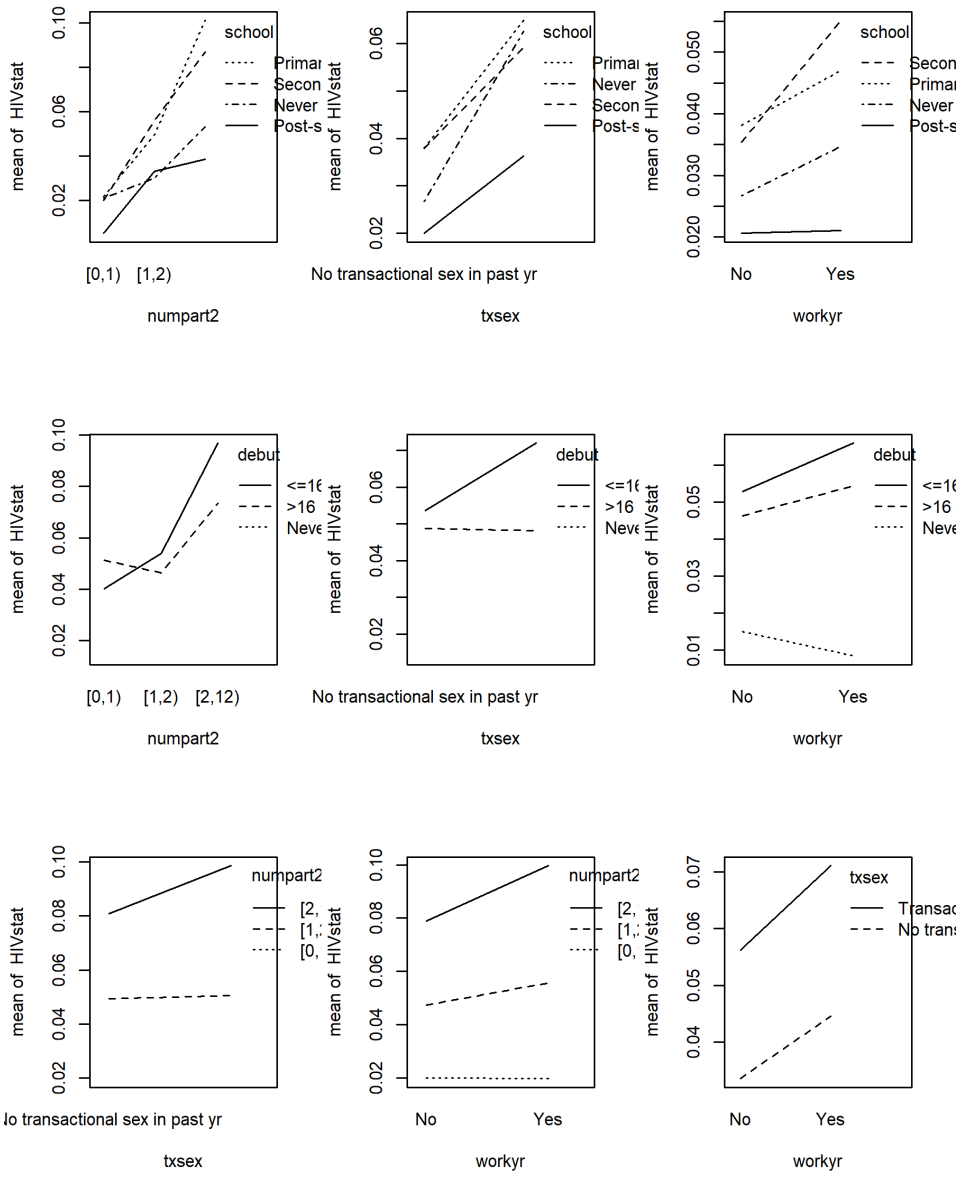

**Supplemental Figure 1.** Interaction plots part 1 of 4.




**Supplemental Figure 2.** Interaction plots part 2 of 4.

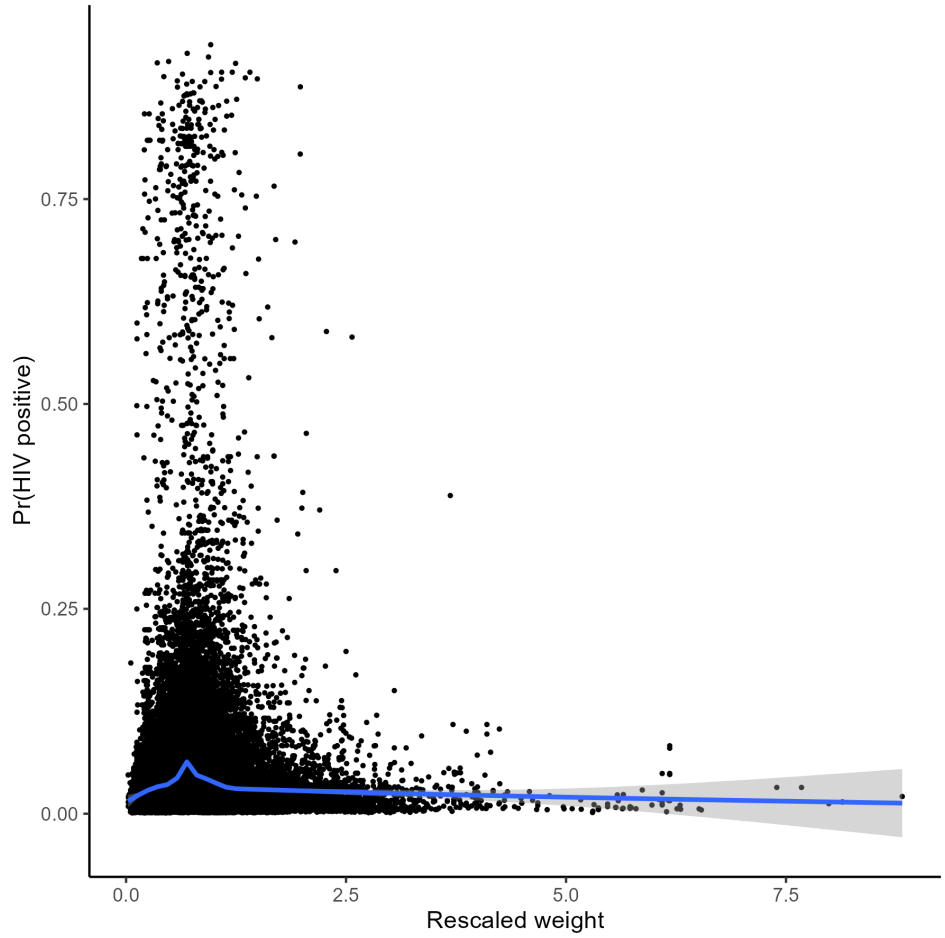



**Supplemental Figure 3.** Interaction plots part 3 of 4.



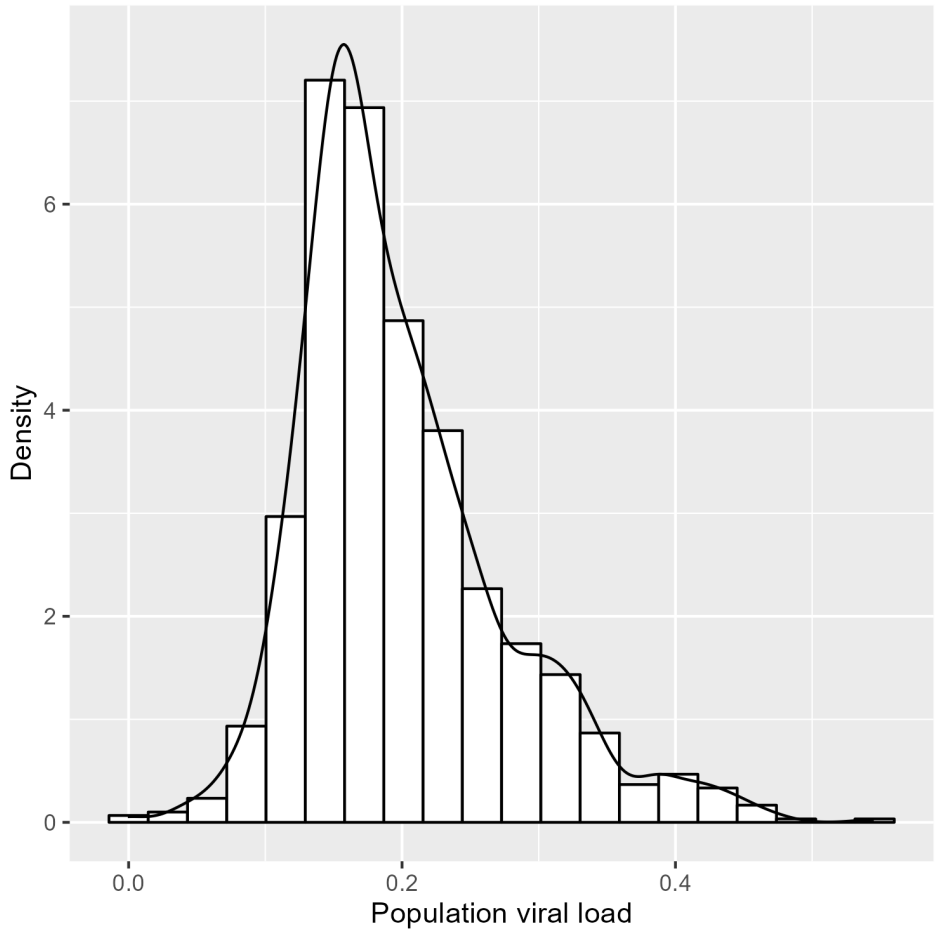

**Supplemental Figure 4.** Interaction plots part 4 of 4.

## 2.2 Rescaled survey weights

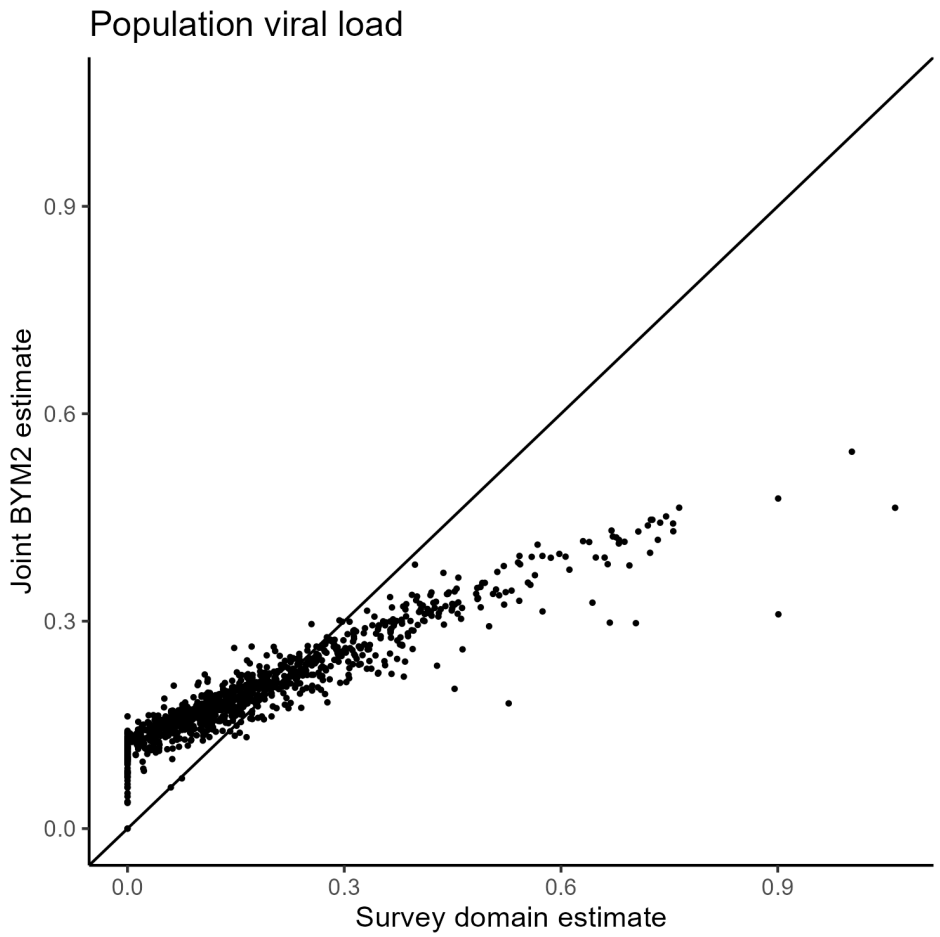



**Supplemental Figure 5.** Predicted probability of testing HIV positive among adolescent girls and young women versus rescaled survey weights from 13 sub-Saharan African countries.

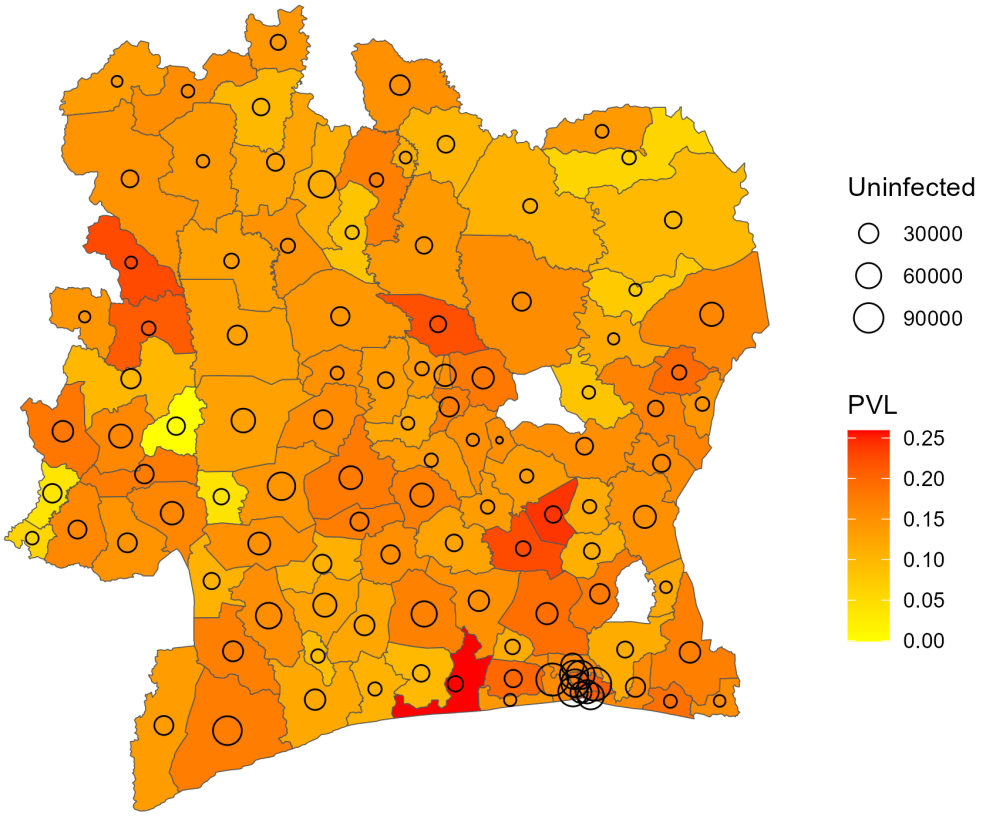
## 2.3 Population viral load



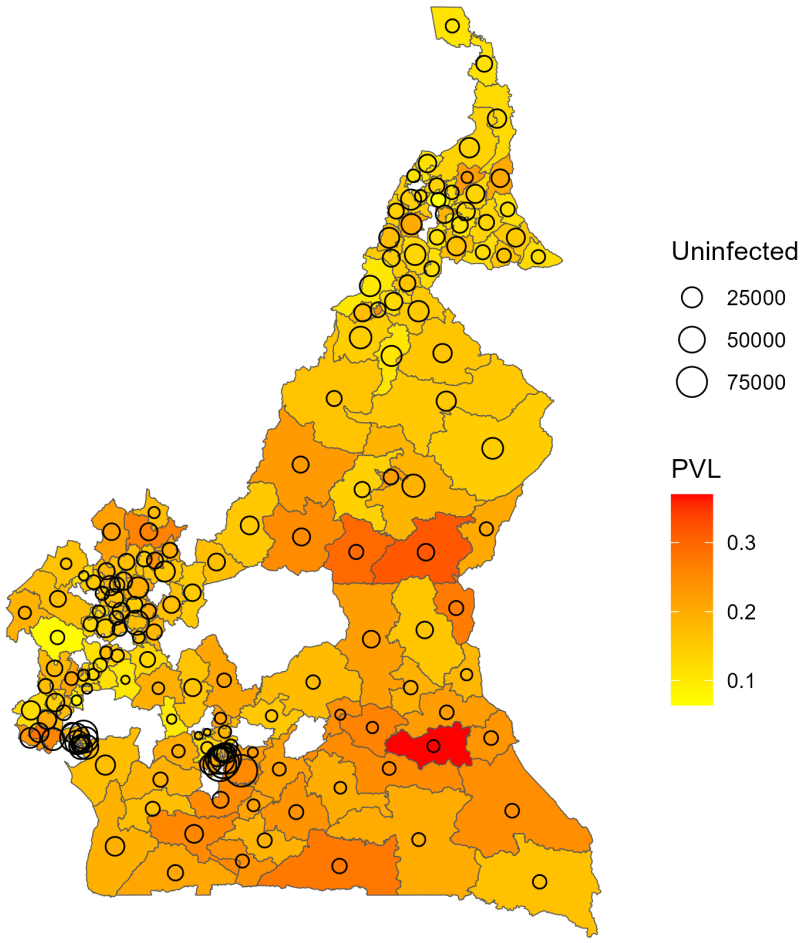

**Supplemental Figure 6.** Marginal posterior distribution of population viral load, measured as $\log_{10}\left(\text{copies} \cdot \text{ml}^{-1} + 1\right)$ from among all PHIA survey respondents including HIV-negative respondents who, by defintion, have a viral load of zero.



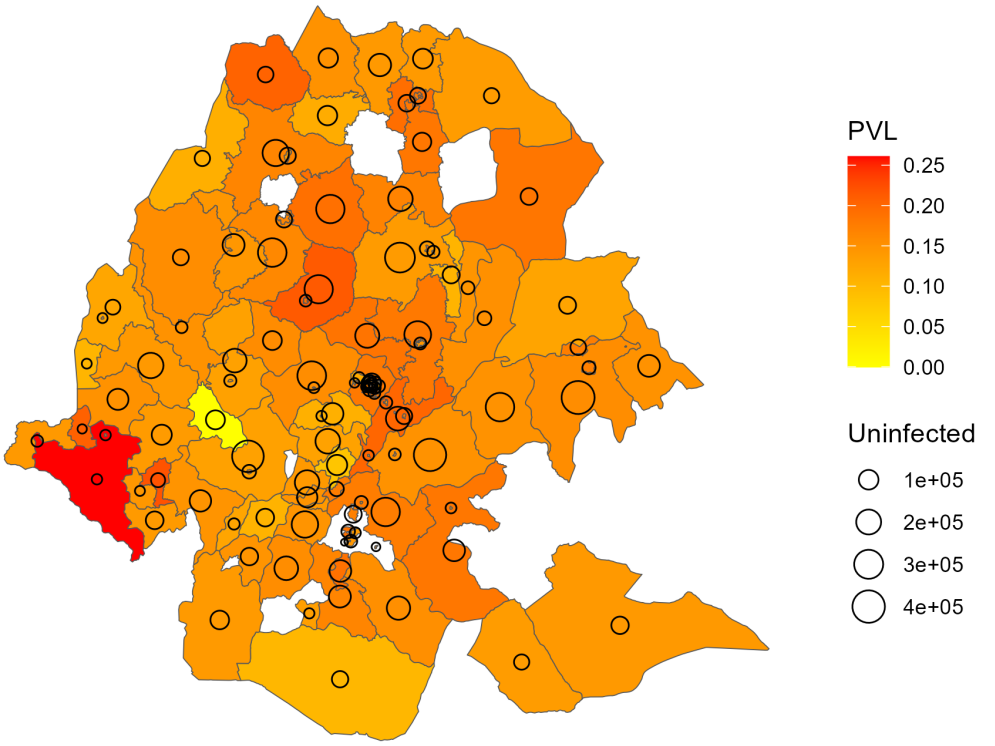

**Supplemental Figure 7.** Relation between spatially smoothed joint estimates and survey domain estimates estimates of population viral load, measured as $\log_{10}\left(\text{copies} \cdot \text{ml}^{-1} + 1\right)$ from among all PHIA survey respondents including HIV-negative respondents who, by definition, have a viral load of zero. The PHIA surveys were not designed to estimate viral loads at the second subnational level and are therefore unstable. The BYM2 spatial smoothing produced considerable shrinkage toward the mean, as expected.



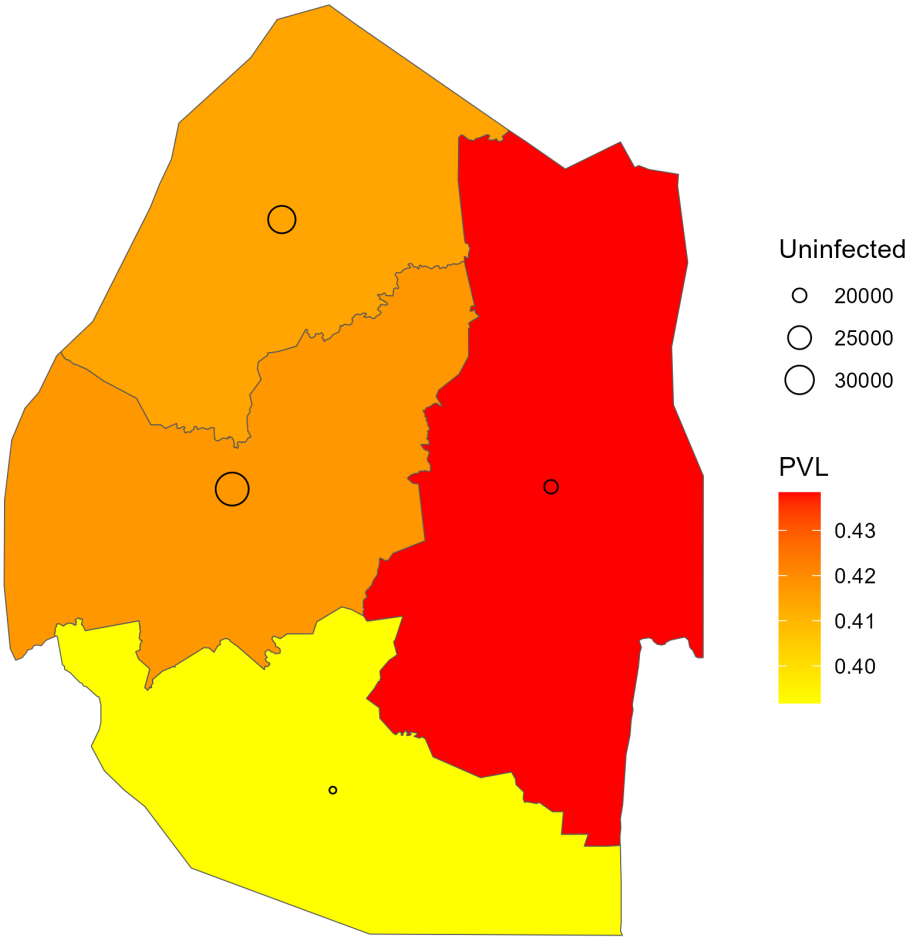

**Supplemental Figure 8.** Estimates of population viral load (PVL; $\log_{10}\left(\text{copies} \cdot \text{ml}^{-1} + 1\right)$) and census-based numbers of HIV-negative adolescent girls and young women of ages 15–24 projected to 2022, Cote d'Ivoire.



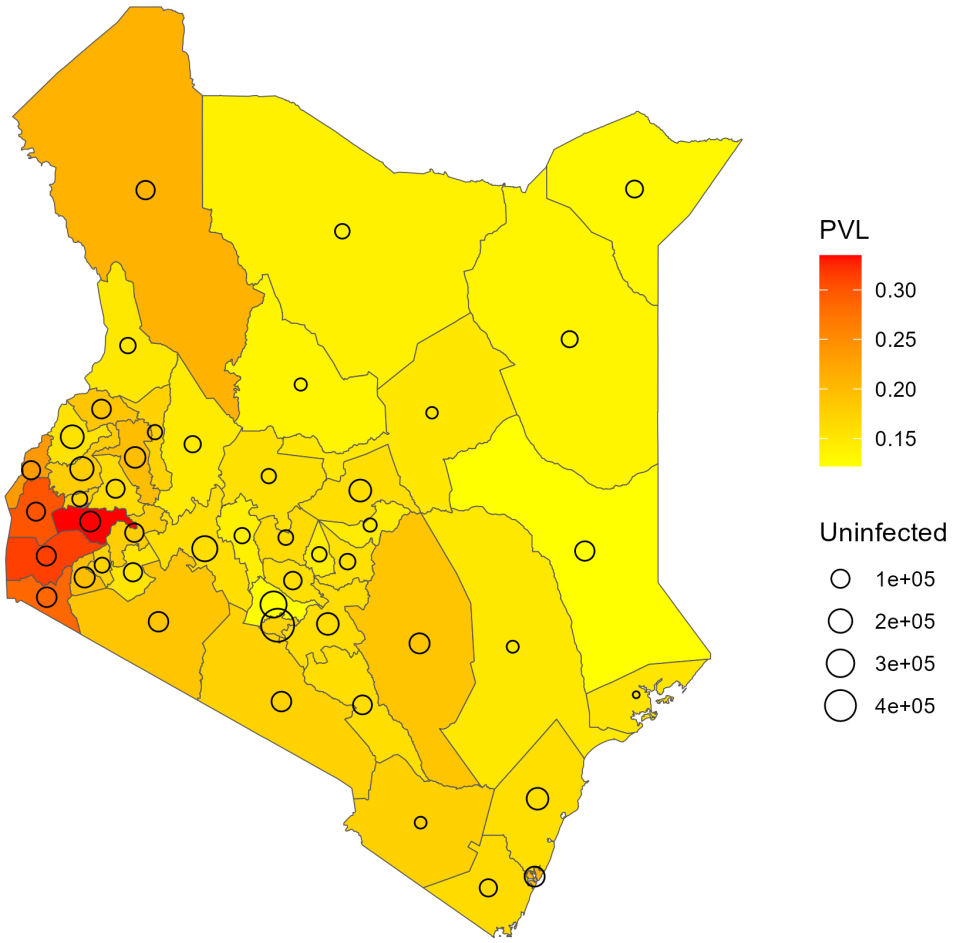

**Supplemental Figure 9.** Estimates of population viral load (PVL; $\log_{10}\left(\text{copies} \cdot \text{ml}^{-1} + 1\right)$) and census-based numbers of HIV-negative adolescent girls and young women of ages 15–24 projected to 2022, Cameroon.



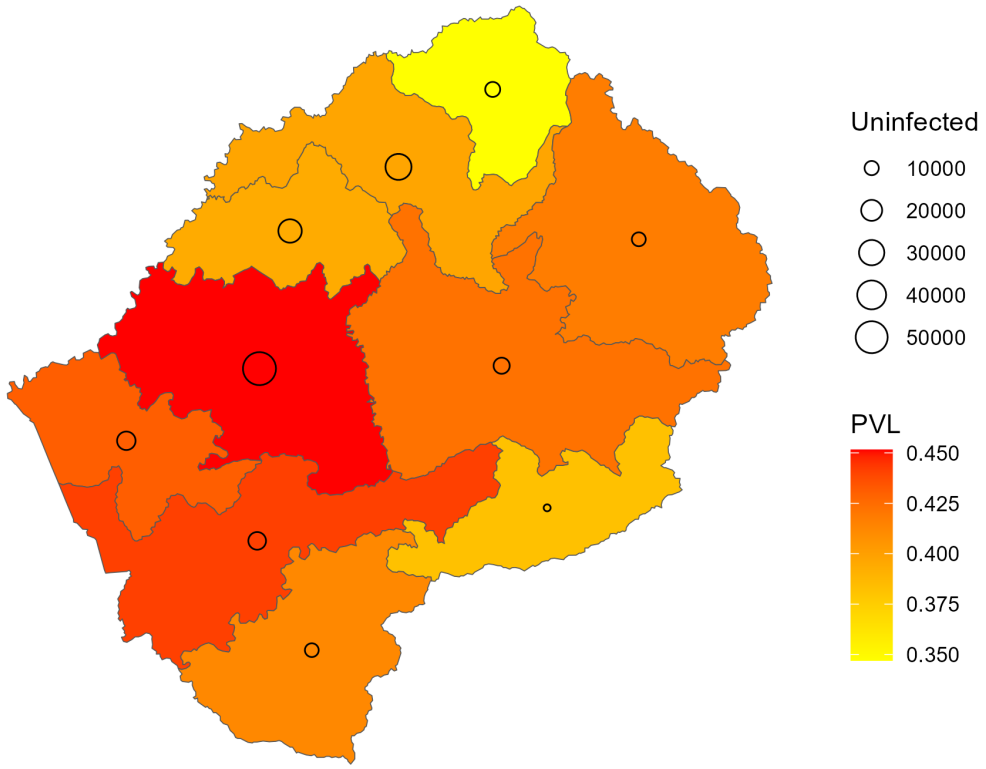

**Supplemental Figure 10.** Estimates of population viral load (PVL; $\log_{10}\left(\text{copies} \cdot \text{ml}^{-1} + 1\right)$) and census-based numbers of HIV-negative adolescent girls and young women of ages 15–24 projected to 2022, Ethiopia.



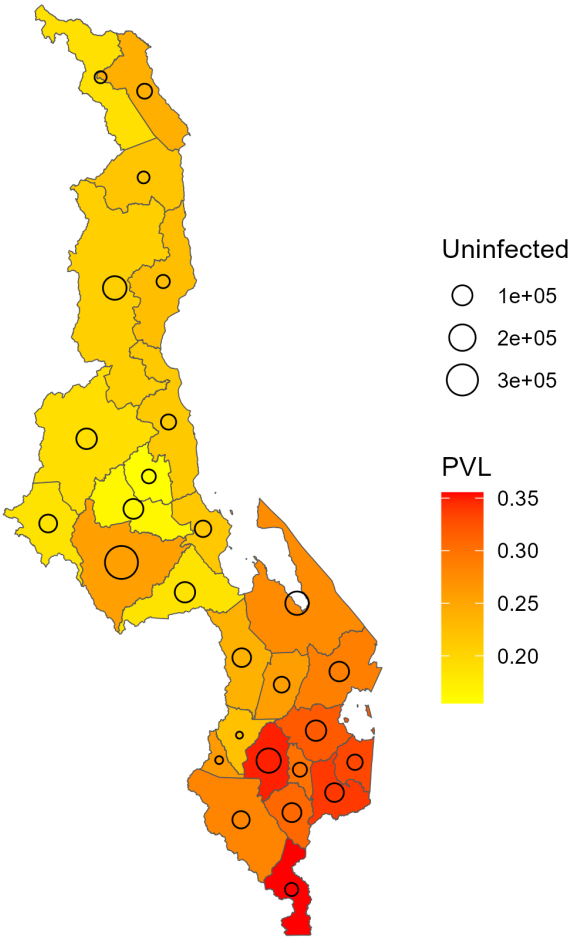

**Supplemental Figure 11.** Estimates of population viral load (PVL; $\log_{10}(\text{copies} \cdot \text{ml}^{-1} + 1)$) and census-based numbers of HIV-negative adolescent girls and young women of ages 15–24 projected to 2022, Eswatini.



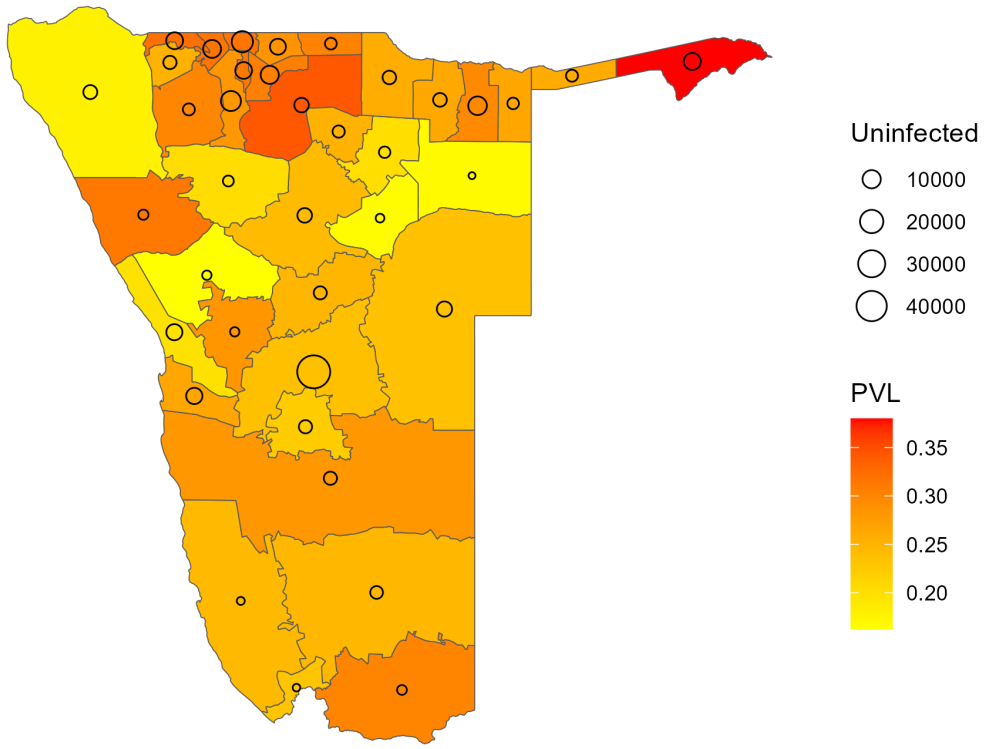

**Supplemental Figure 12.** Estimates of population viral load (PVL; $\log_{10}(\text{copies} \cdot \text{ml}^{-1} + 1)$) and census-based numbers of HIV-negative adolescent girls and young women of ages 15–24 projected to 2022, Kenya.



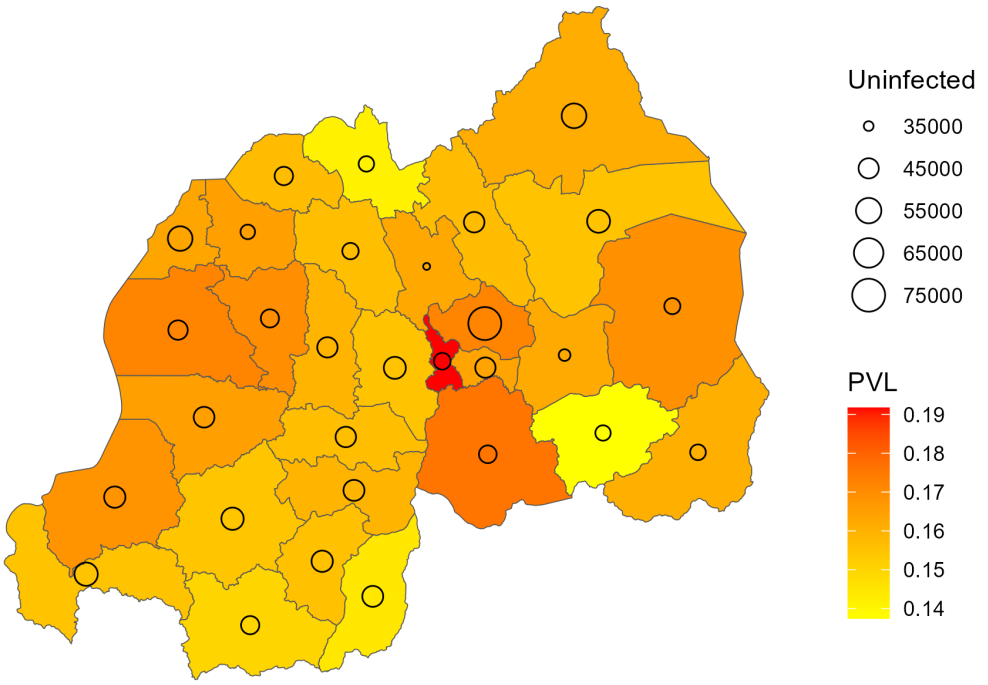

**Supplemental Figure 13.** Estimates of population viral load (PVL; $\log_{10}\left(\text{copies} \cdot \text{ml}^{-1} + 1\right)$) and census-based numbers of HIV-negative adolescent girls and young women of ages 15–24 projected to 2022, Lesotho.



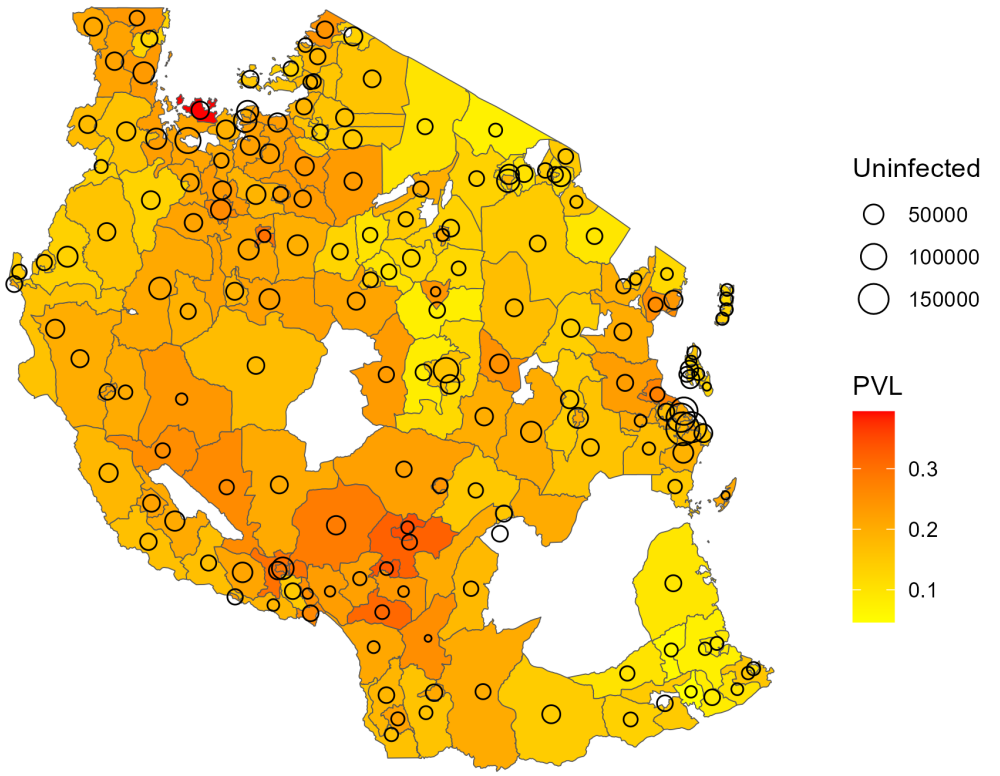

**Supplemental Figure 14.** Estimates of population viral load (PVL; $\log_{10}(\text{copies} \cdot \text{ml}^{-1} + 1)$) and census-based numbers of HIV-negative adolescent girls and young women of ages 15–24 projected to 2022, Malawi.



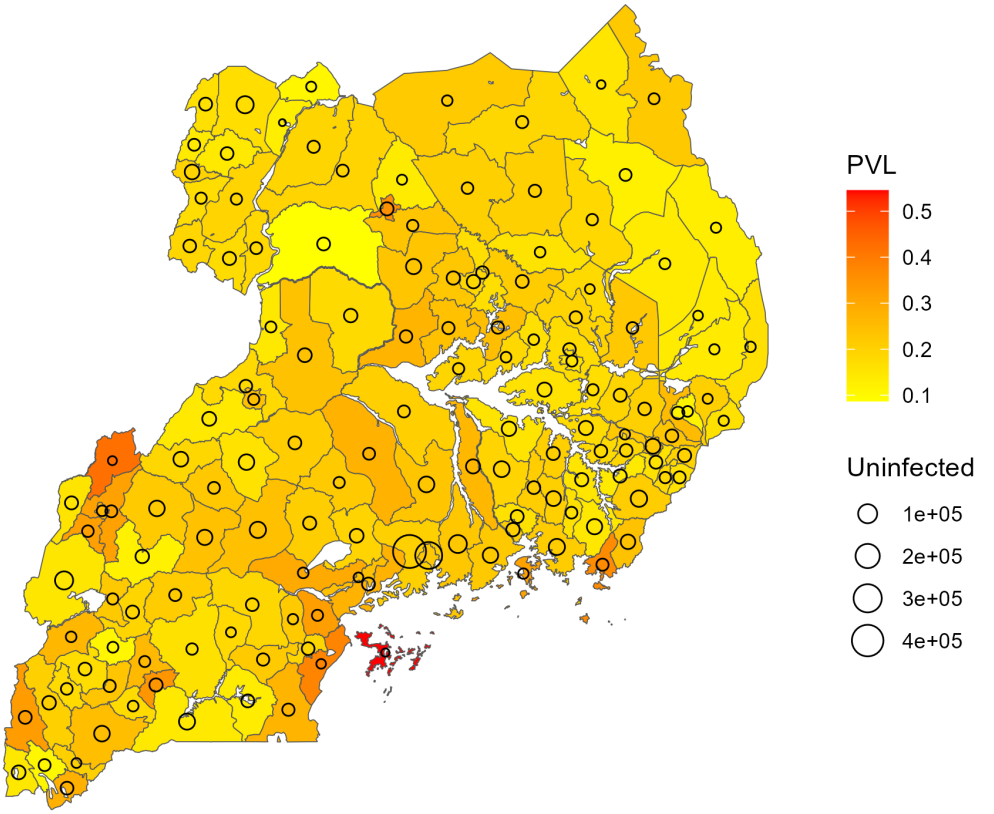

**Supplemental Figure 15.** Estimates of population viral load (PVL; $\log_{10}\left(\text{copies} \cdot \text{ml}^{-1} + 1\right)$) and census-based numbers of HIV-negative adolescent girls and young women of ages 15–24 projected to 2022, Namibia.



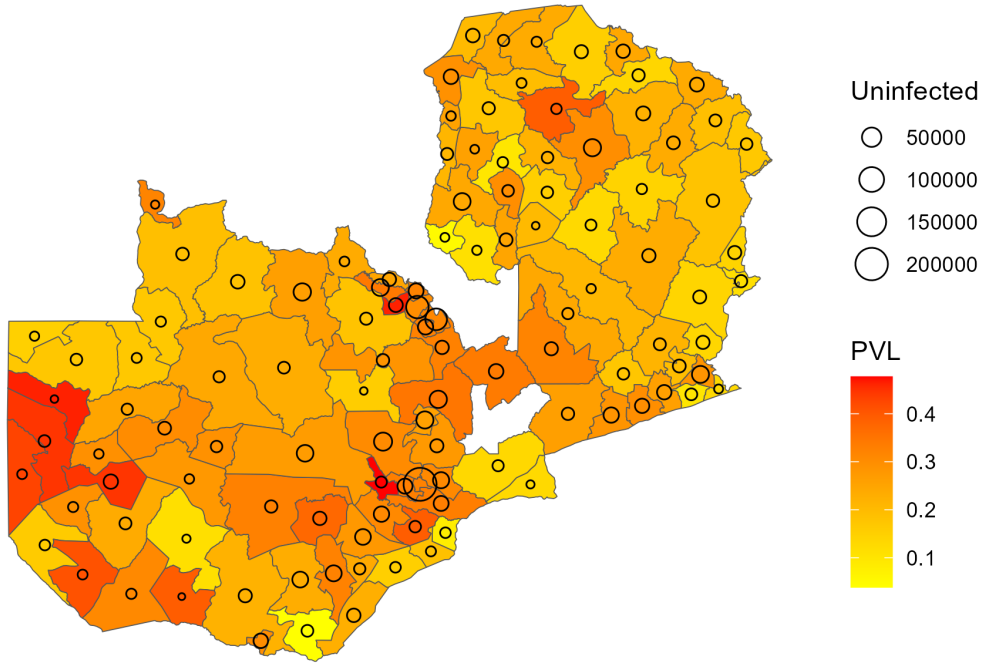

**Supplemental Figure 16.** Estimates of population viral load (PVL; $\log_{10}(\text{copies} \cdot \text{ml}^{-1} + 1)$) and census-based numbers of HIV-negative adolescent girls and young women of ages 15–24 projected to 2022, Rwanda.



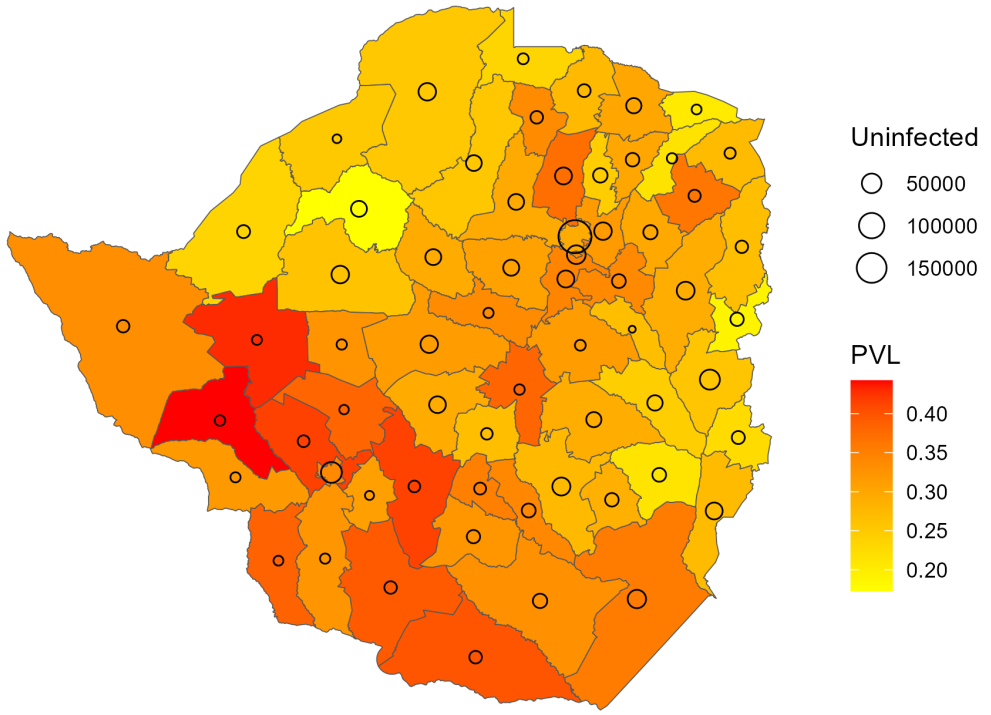

**Supplemental Figure 17.** Estimates of population viral load (PVL; $\log_{10}(\text{copies} \cdot \text{ml}^{-1} + 1)$) and census-based numbers of HIV-negative adolescent girls and young women of ages 15–24 projected to 2022, Tanzania.




**Supplemental Figure 18.** Estimates of population viral load (PVL; $\log_{10}\left(\text{copies} \cdot \text{ml}^{-1} + 1\right)$) and census-based numbers of HIV-negative adolescent girls and young women of ages 15–24 projected to 2022, Uganda.

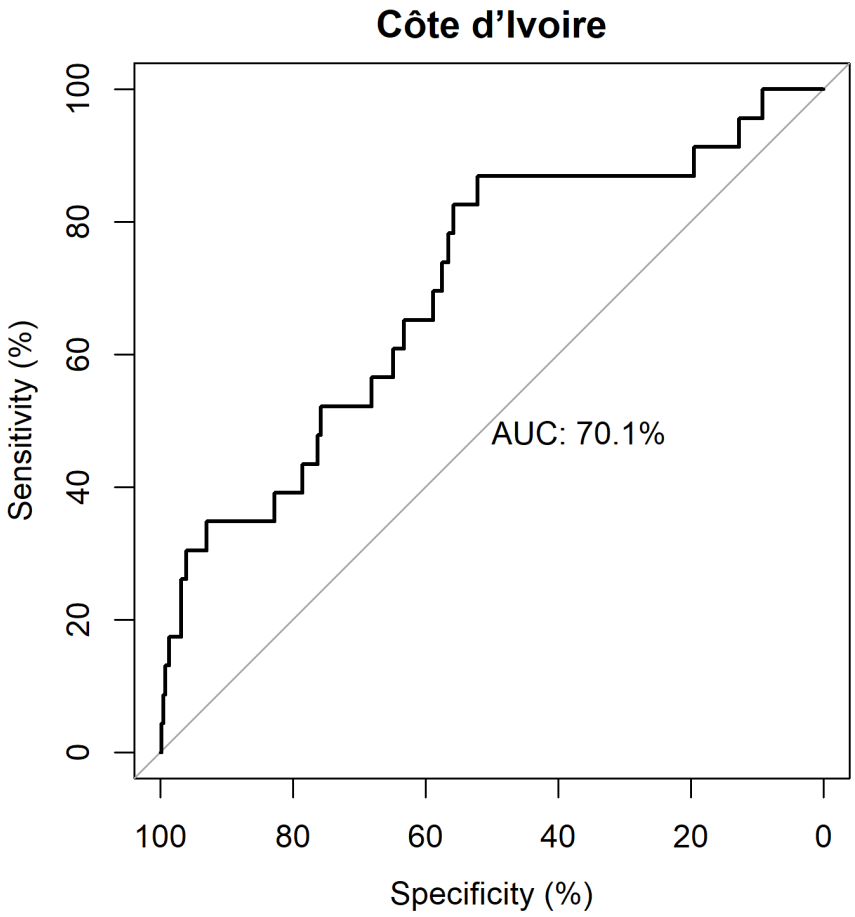



**Supplemental Figure 19.** Estimates of population viral load (PVL; $\log_{10}(\text{copies} \cdot \text{ml}^{-1} + 1)$) and census-based numbers of HIV-negative adolescent girls and young women of ages 15–24 projected to 2022, Zambia.

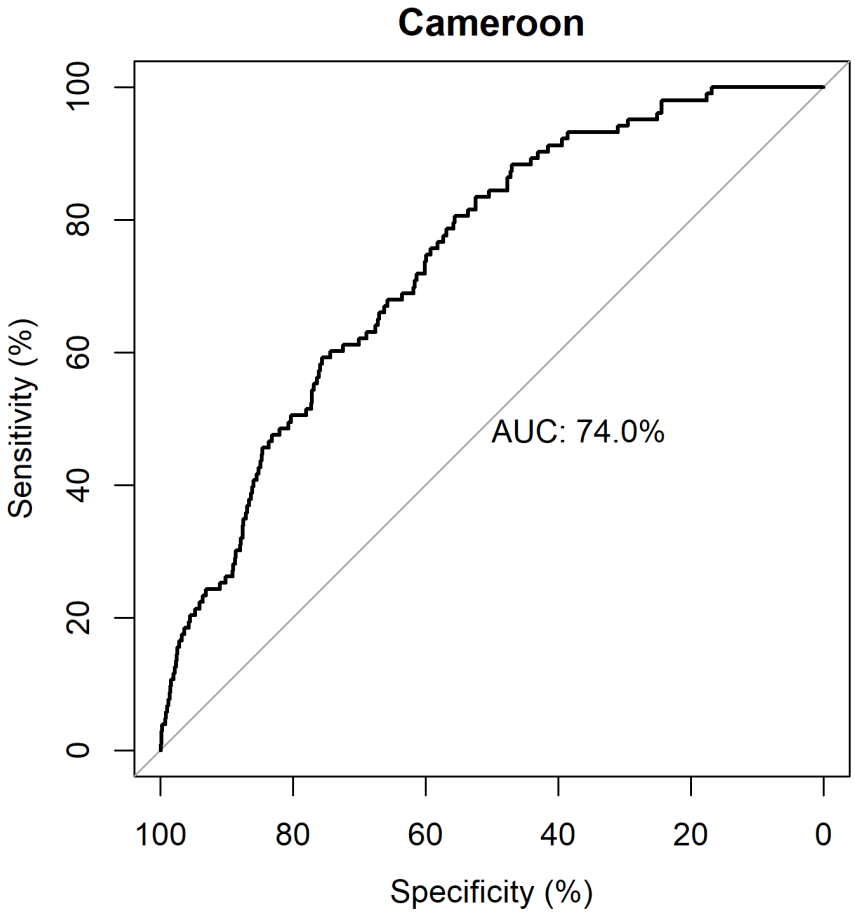



**Supplemental Figure 20.** Estimates of population viral load (PVL; $\log_{10}\left(\text{copies} \cdot \text{ml}^{-1} + 1\right)$) and census-based numbers of HIV-negative adolescent girls and young women of ages 15–24 projected to 2022, Zimbabwe.

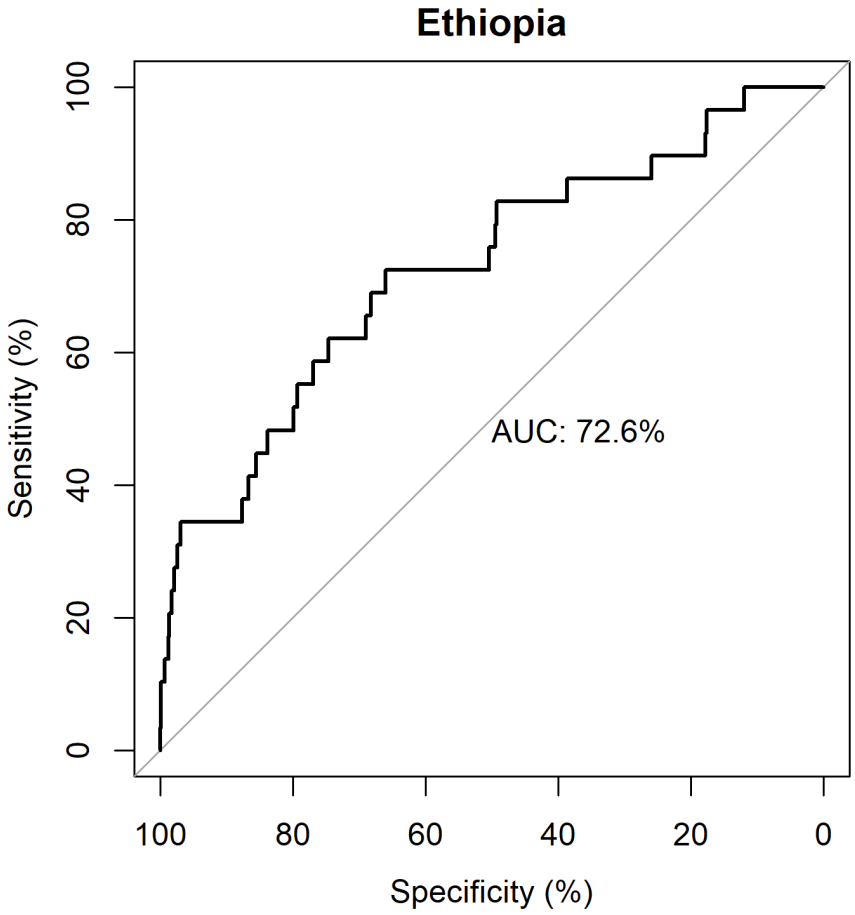
## 2.4 Country-specific ROC curves



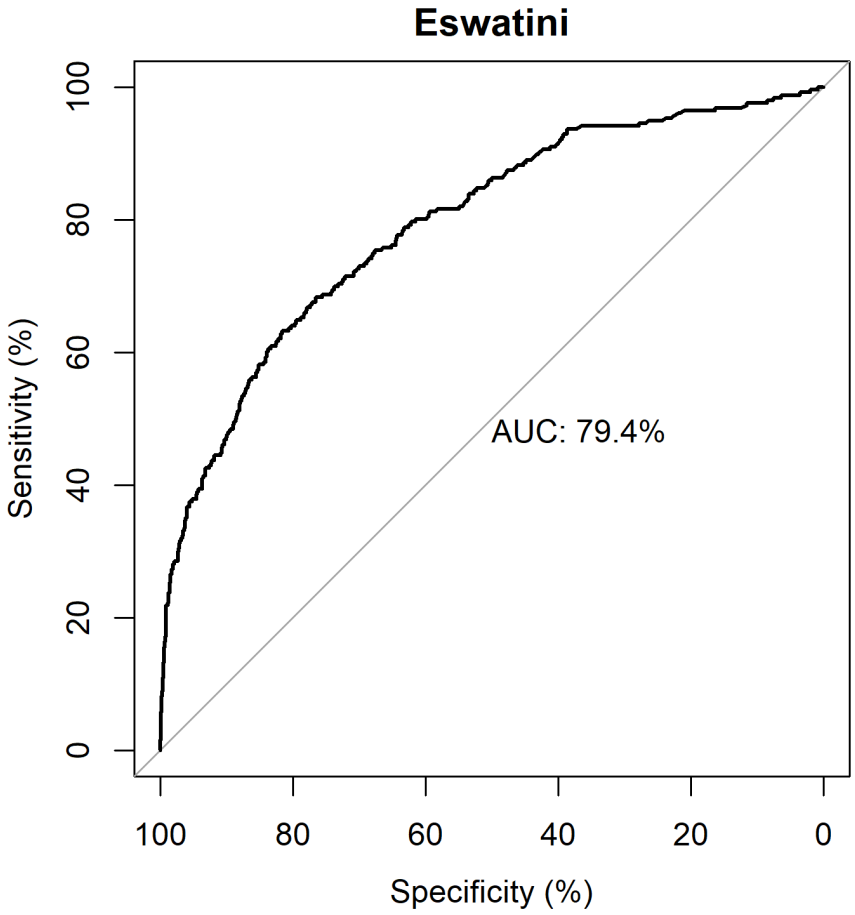

**Supplemental Figure 21.** Estimated out-of-sample performance of the joint model for classification of HIV status among adolescent girls and young women of ages 15–24 in Cote d'Ivoire, as measured by the receiver-operating characteristic curve. The area under the curve is denoted by AUC.



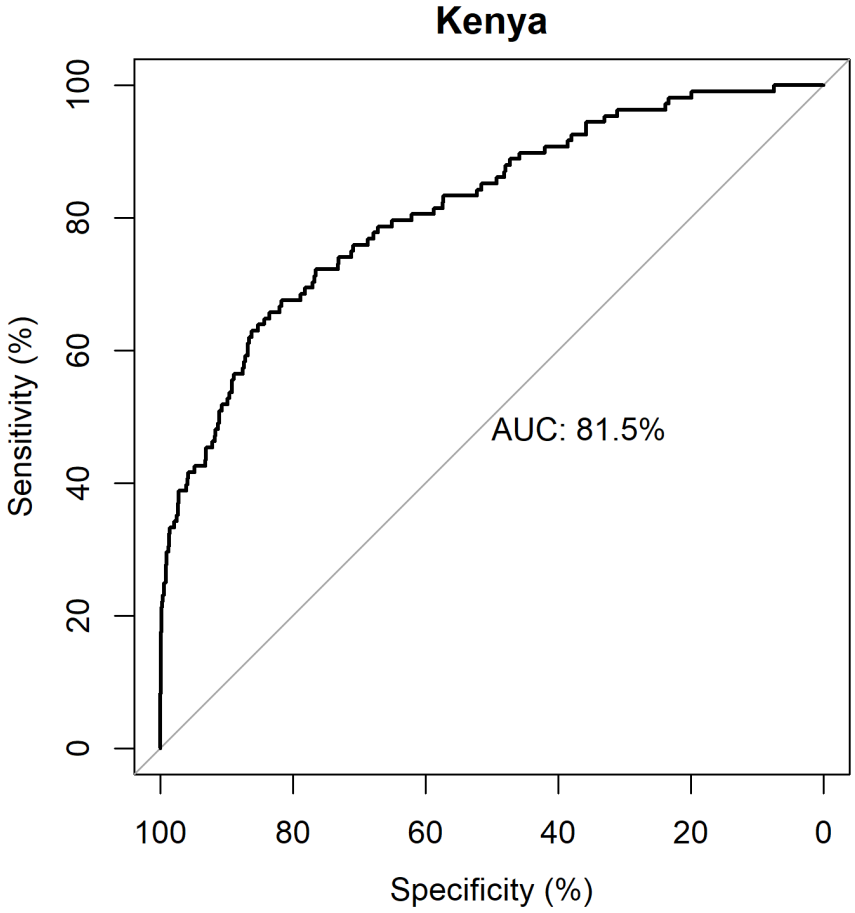

**Supplemental Figure 22.** Estimated out-of-sample performance of the joint model for classification of HIV status among adolescent girls and young women of ages 15–24 in Cameroon, as measured by the receiver-operating characteristic curve. The area under the curve is denoted by AUC.



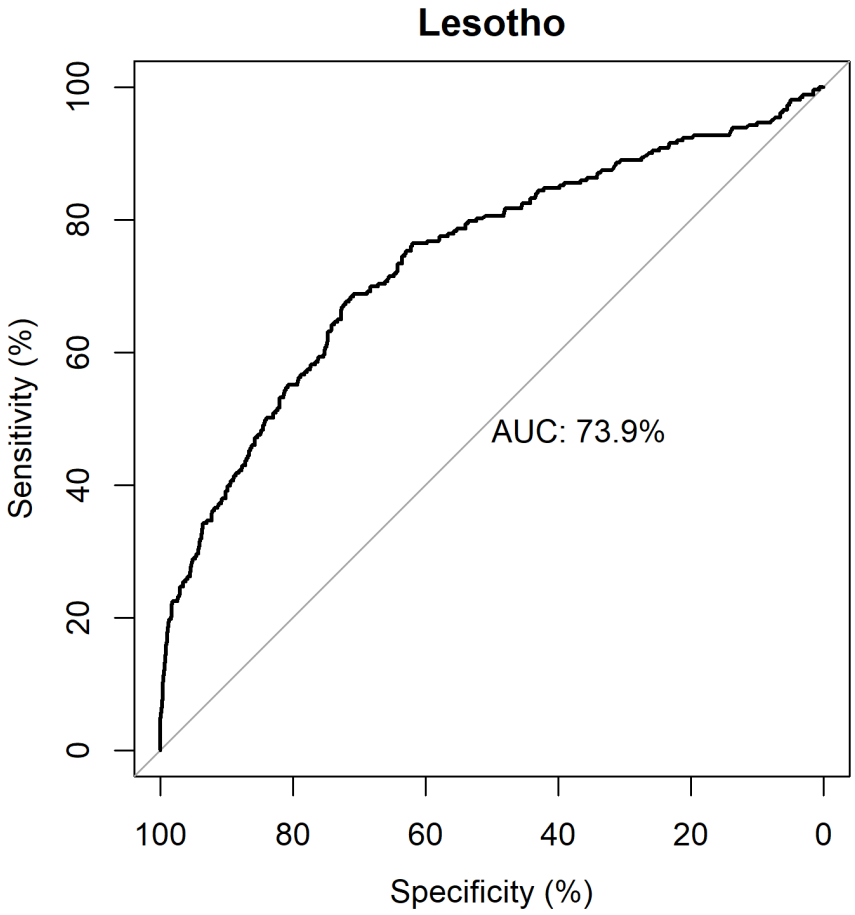

**Supplemental Figure 23.** Estimated out-of-sample performance of the joint model for classification of HIV status among adolescent girls and young women of ages 15–24 in Ethiopia, as measured by the receiver-operating characteristic curve. The area under the curve is denoted by AUC.



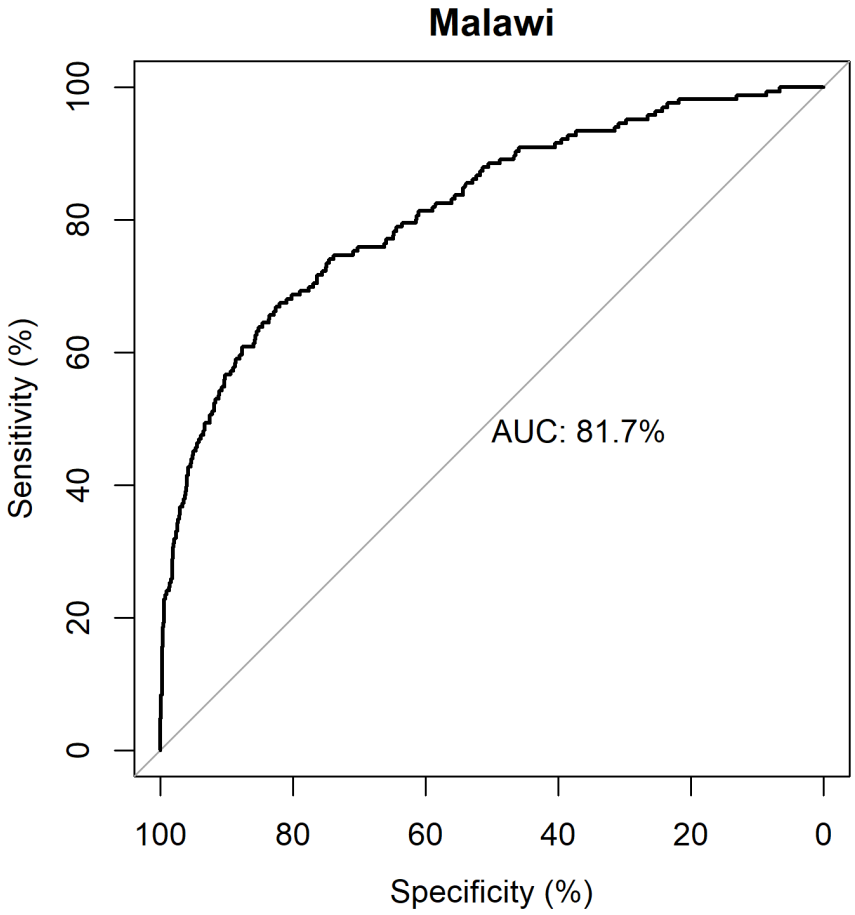

**Supplemental Figure 24.** Estimated out-of-sample performance of the joint model for classification of HIV status among adolescent girls and young women of ages 15–24 in Essatini, as measured by the receiver-operating characteristic curve. The area under the curve is denoted by AUC.



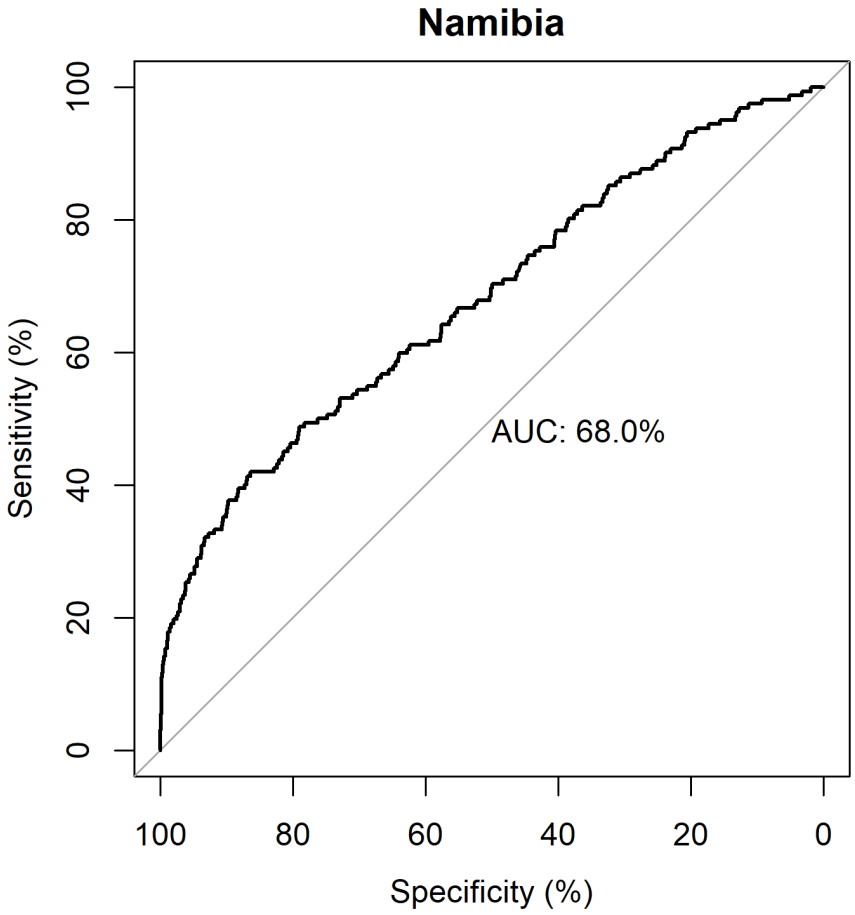

**Supplemental Figure 25.** Estimated out-of-sample performance of the joint model for classification of HIV status among adolescent girls and young women of ages 15–24 in Kenya, as measured by the receiver-operating characteristic curve. The area under the curve is denoted by AUC.



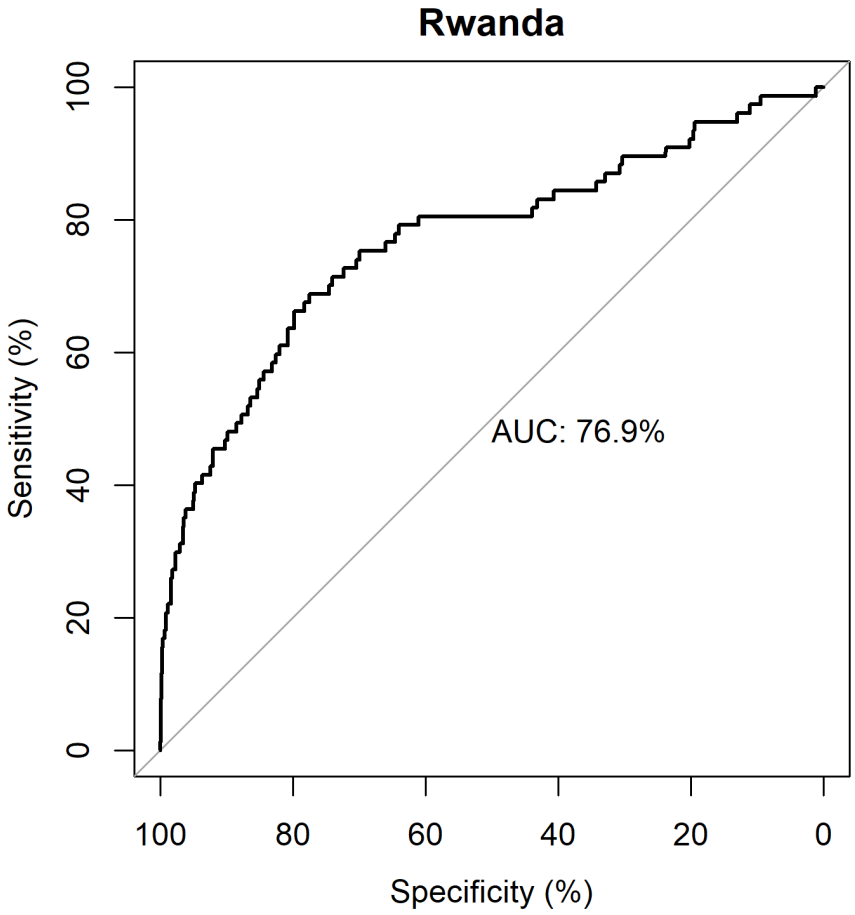

**Supplemental Figure 26.** Estimated out-of-sample performance of the joint model for classification of HIV status among adolescent girls and young women of ages 15–24 in Lesotho, as measured by the receiver-operating characteristic curve. The area under the curve is denoted by AUC.



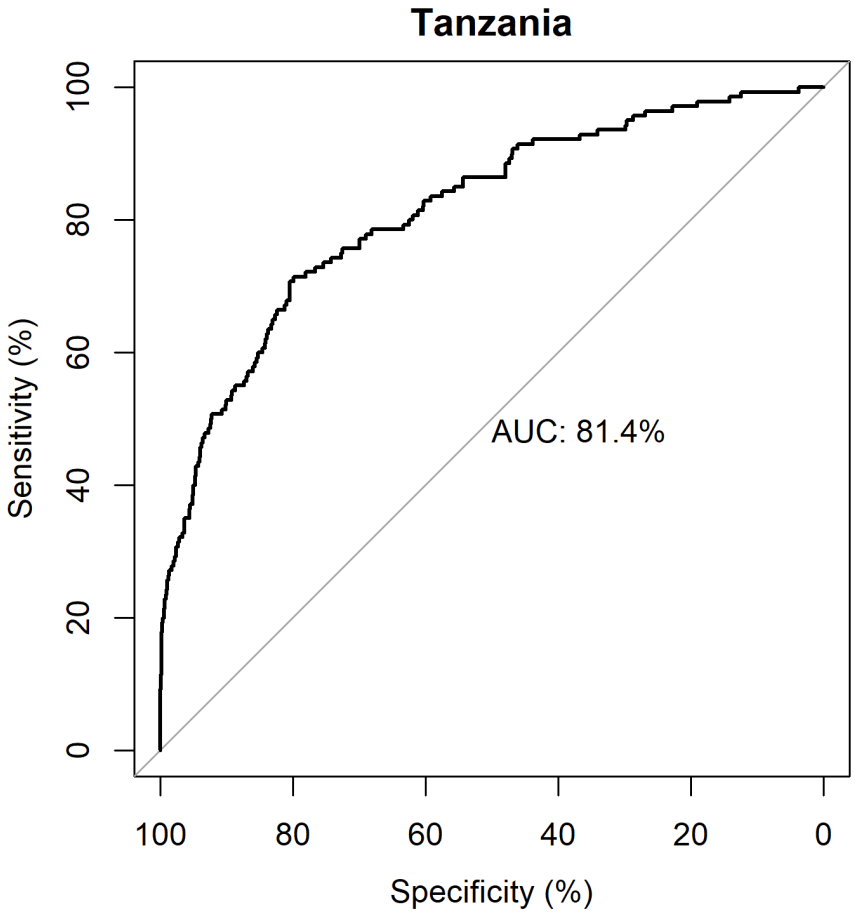

**Supplemental Figure 27.** Estimated out-of-sample performance of the joint model for classification of HIV status among adolescent girls and young women of ages 15–24 in Malawi, as measured by the receiver-operating characteristic curve. The area under the curve is denoted by AUC.



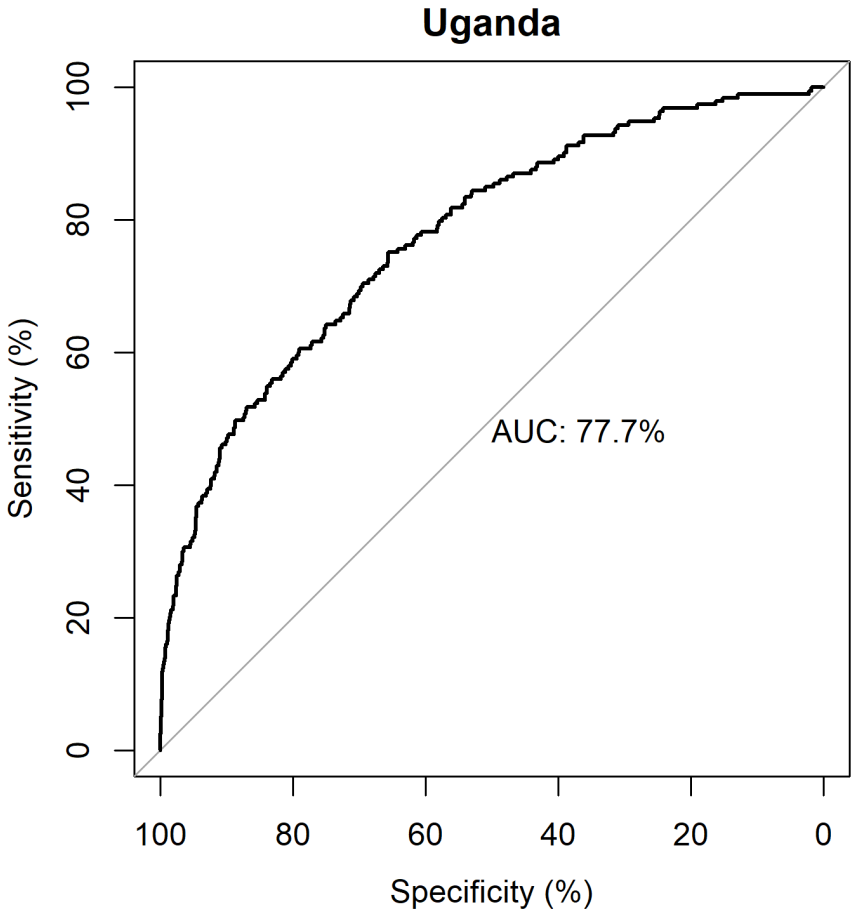

**Supplemental Figure 28.** Estimated out-of-sample performance of the joint model for classification of HIV status among adolescent girls and young women of ages 15–24 in Namibia, as measured by the receiver-operating characteristic curve. The area under the curve is denoted by AUC.



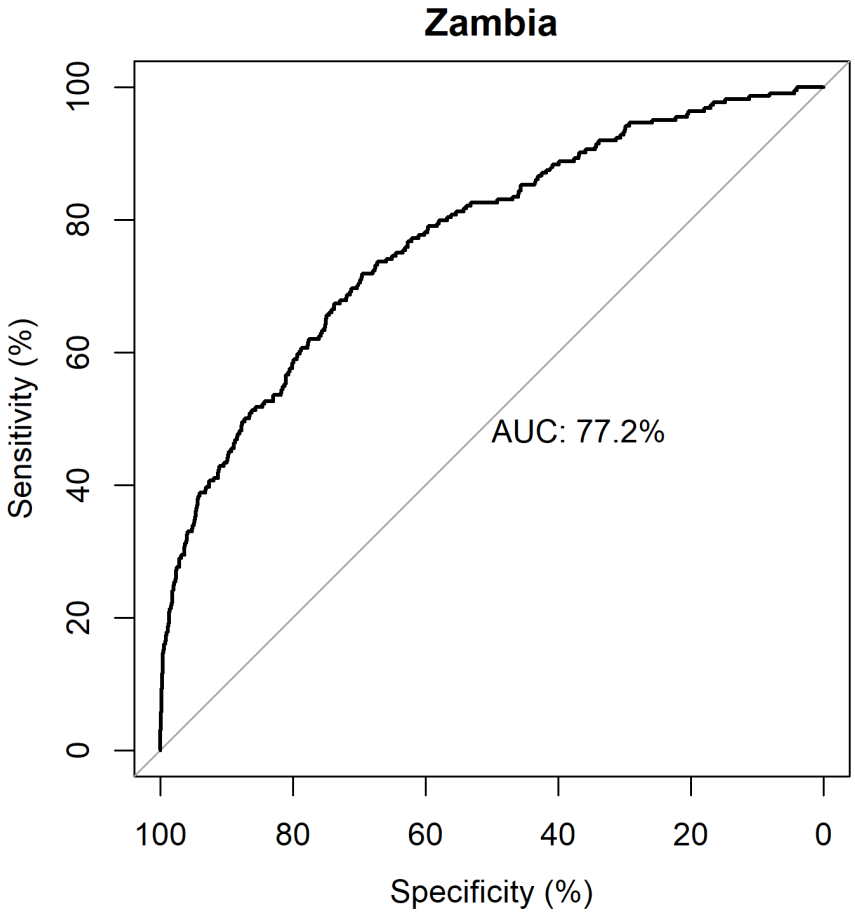

**Supplemental Figure 29.** Estimated out-of-sample performance of the joint model for classification of HIV status among adolescent girls and young women of ages 15–24 in Rwanda, as measured by the receiver-operating characteristic curve. The area under the curve is denoted by AUC.



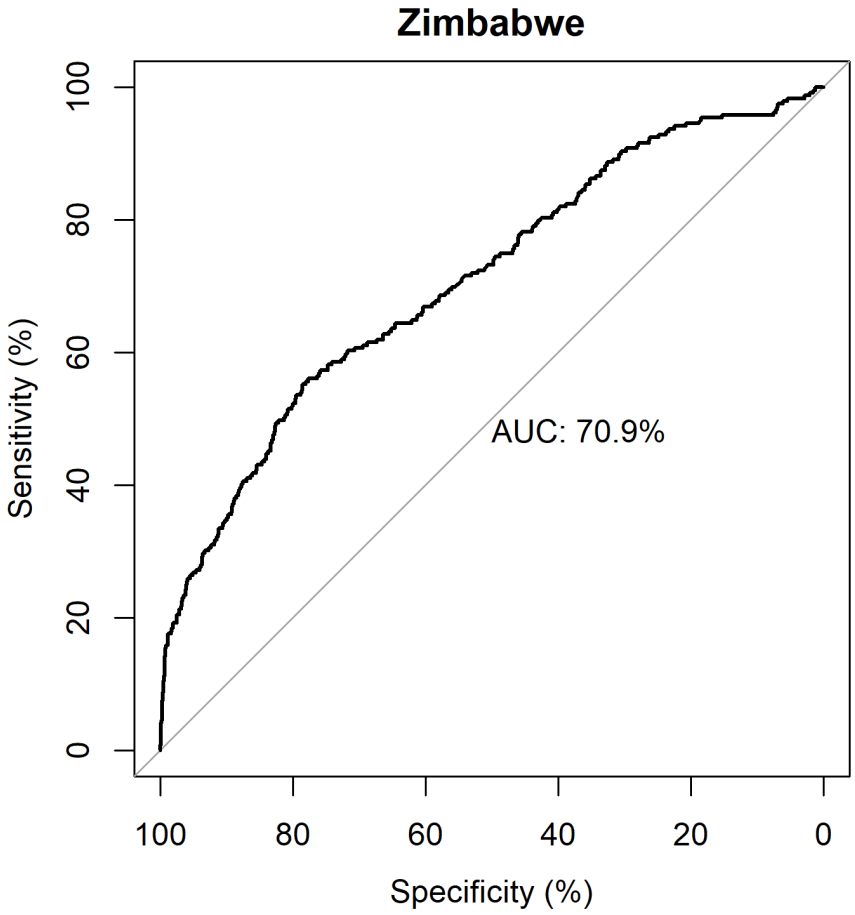

**Supplemental Figure 30.** Estimated out-of-sample performance of the joint model for classification of HIV status among adolescent girls and young women of ages 15–24 in Tanzania, as measured by the receiver-operating characteristic curve. The area under the curve is denoted by AUC.




**Supplemental Figure 31.** Estimated out-of-sample performance of the joint model for classification of HIV status among adolescent girls and young women of ages 15–24 in Uganda, as measured by the receiver-operating characteristic curve. The area under the curve is denoted by AUC.




**Supplemental Figure 32.** Estimated out-of-sample performance of the joint model for classification of HIV status among adolescent girls and young women of ages 15–24 in Zambia, as measured by the receiver-operating characteristic curve. The area under the curve is denoted by AUC.




**Supplemental Figure 33.** Estimated out-of-sample performance of the joint model for classification of HIV status among adolescent girls and young women of ages 15–24 in Zimbabwe, as measured by the receiver-operating characteristic curve. The area under the curve is denoted by AUC.

# 3 Plug-in Prediction from New Observations

Posterior prediction of new observations using INLA requrires augmenting the data with new rows of predictors and random effects for the newly encountered AGYW and then refitting a model. INLA then predicts the outcome (HIV status) for the augmented rows of data. That approach is impractical given the computational burden of model fitting. Instead, we propose simple plug-in prediction of the probability of testing positive among newly encountered AGYW.

The linear predictor for the final model (M100; Subsection 1.2.1) is given by

$$\begin{aligned}
\text{logit}(p_{mij}) = &-6.3394 + 8.5141\text{pvl}_j + 0.5007\text{urban}_{mij} \\
&+ 0.0024\text{debut1_age0_workyr0}_{mij} + 0.3502\text{debut2_age0_workyr0}_{mij} \\
&- 0.2814\text{debut0_age1}_{mij} + 0.7762\text{debut1_age1}_{mij} \\
&+ 1.0781\text{debut2_age1}_{mij} - 0.2994\text{debut0_workyr1}_{mij} \\
&+ 0.2356\text{debut1_workyr1}_{mij} + 0.2594\text{debut2_workyr1}_{mij} \\
&- 0.1447\text{preg0_partN}_{mij} + 0.3579\text{preg0_partU}_{mij} \\
&+ 2.4949\text{preg0_partP}_m ij + 0.5759\text{preg1_part0}_{mij} \\
&- 2.310\text{preg1_partN}_m ij + 0.6846\text{preg1_partU}_{mij} \\
&+ 3.4677\text{preg1_partP}_{mij} - 0.0472\text{schPri}_{mij} \\
&+ 0.3836\text{schSec}_{mij} - 1.0479\text{schPSec}_{mij} \\
&+ 0.2404\text{txsex} + \text{v_c}_m + \text{b_y}_j
\end{aligned}$$

for newly encountered AGYW $i$ from country $m$ and TSNU $j$, where the predictors are defined in Supplemental Table 3, the $\text{v\_c}_m$ are the country-level random effect estimates (Supplemental Digital Content 2, M100_country_re_estimates.csv), and the $\text{pvl}_j$ and $\text{b\_y}_j$ are the TSNU-level PVL and BYM2 random effect estimates, respectively (Supplemental Digital Content 2, M100_BYM2_PVL+b_y_estimates.csv). Then the predicted probability of HIV infection is given by $\text{expit}\left[\text{logit}\left(p_{mij}\right)\right]$.

**Supplemental Table 3.** Predictor definitions.

| Predictor | Definition |
|---|---|
| urban | 1 if urban residency; 0 otherwise |
| debut1_age0_workyr0 | 1 if sexual debut $\leq$ 16 and age 15–19 and no employment; 0 otherwise |
| debut2_age0_workyr0 | 1 if sexual debut $>$ 16 and age 15–19 and no employment; 0 otherwise |
| debut0_age1 | 1 if never had sex and age 20–24; 0 otherwise |
| debut1_age1 | 1 if sexual debut $\leq$ 16 and age 20–24; 0 otherwise |
| debut2_age1 | 1 if sexual debut  16 and age 20–24; 0 otherwise |
| debut0_workyr1 | 1 if never had sex and worked during past year; 0 otherwise |
| debut1_workyr1 | 1 if sexual debut  16 and worked; 0 otherwise |
| debut2_workyr1 | 1 if sexual debut $\leq$ 16 and worked; 0 otherwise |
| preg0_partN | 1 if never pregnant and partner HIV negative; 0 otherwise |
| preg0_partU | 1 if never pregnant and partner status unknown; 0 otherwise |
| preg0_partP | 1 if never pregnant and partner HIV positive; 0 otherwise |
| preg1_part0 | 1 if ever pregnant and no current partner; 0 otherwise |
| preg1_partN | 1 if ever pregnant and partner HIV negative; 0 otherwise |
| preg1_partU | 1 if ever pregnant and partner status unknown; 0 otherwise |
| preg1_partP | 1 if never pregnant and partner HIV positive; 0 otherwise |
| schPri | 1 if completed only primary education; 0 otherwise |
| schSec | 1 if completed only secondary education; 0 otherwise |
| schPSec | 1 if some post-secondary education; 0 otherwise |
| txsex | 1 if engaged in transactional sex during past year; 0 otherwise |

## References


[1] Little RJ. To model or not to model? Competing modes of inference for finite population sampling. Journal of the American Statistical Association. 2004;99(466):546-56. doi:10.1198/016214504000000467.

[2] Gelman A. Struggles with survey weighting and regression modeling. Statistical Science. 2007;22:153-64. doi:10.1214/088342306000000691.

[3] Wakefield J, Okonek T, Pedersen J. Small area estimation for disease prevalence mapping. International Statistical Review. 2020. doi:10.1111/insr.12400.

[4] Paige J, Fuglstad GA, Riebler A, Wakefield J. Design- and model-based approaches to small-area estimation in a low- and middle-income country context: Comparisons and recommendations. Journal of Statistics and Survey Methodology. 2022;10(1):50-80. doi:10.1093/jssam/smaa011.

[5] Zheng H, Little RJA. Penalized spline model-based estimation of the finite populations total from probability-proportional-to-size samples. Journal of Official Statistics. 2003;19(2):99-117.

[6] Vandendijck Y, Faes C, Hens N. Prevalence and trend estimation from observational data with highly variable post-stratification weights. Annals of Applied Statistics. 2016;10(1):94-17. doi:10.1214/15-AOAS874.

[7] Chen Q, Elliott MR, Little RJA. Bayesian penalized spline model-based inference for finite population proportion in unequal probability sampling. Survey Methodology. 2010;36(1):23-34.

[8] Pfeffermann D, Skinner C, Holmes D, Goldstein H, Rasbash J. Weighting for unequal selection probabilities in multilevel models. Journal of the Royal Statistical Society Series B. 1998;60(1):23-40. doi:10.1111/1467-9868.00106.

[9] Rabe-Hesketh S, Skrondal A. Multilevel modeling of complex survey data. Journal of the Royal Statistical Society Series A. 2006;169(4):805-27. doi:10.1111/j.1467-985X.2006.00426.x.

[10] Muff S, Riebler A, Held L, Rue H, Saner P. Bayesian analysis of measurement error models using integrated nested Laplace approximations. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2015;64(2):231-52. doi:10.1111/rssc.12069.

[11] Tobler WR. A computer movie simulating urban growth in the Detroit region. Economic Geography. 1970;46:234-40. doi:10.2307/143141.

[12] Simpson D, Rue H, Riebler A, Martins TG, Sørbye SH. Penalising model component complexity: A principled, practical approach to constructing priors. Statistical Science. 2017;32(1):1-28. doi:10.1214/16-STS576.

[13] Riebler A, Sørbye SH, Simpson D, Rue H. An intuitive Bayesian spatial model for disease mapping that accounts for scaling. Statistical Methods in Medical Research. 2016;25(4):1145-65. doi:10.1177/0962280216660421.

[14] Besag J, York J, Mollié A. Bayesian image restoration, with two applications in spatial statistics. Annals of the Institute of Statistical Mathematics. 1991;43(1):1-20.

[15] Sørbye SH, Rue H. Scaling intrinsic Gaussian Markov random field priors in spatial modelling. Spatial Statistics. 2014;8:39-51. doi:10.1016/j.spasta.2013.06.004.

[16] R Core Team. , editor. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2022. Available from: `http://www.R-project.org/`.

[17] Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society: Series B. 2009;71(2):319-92. doi:DOI: 10.1111/j.1467-9868.2008.00700.x.

[18] Eilers PHC, Marx BD. Flexible smoothing with $B$-splines and penalties. Statistical Science. 1996;11(2):89-121. doi:10.1214/ss/1038425655.